Modeling individual and collective opinion in online social networks: drivers of choice behavior and effects of marketing interventions

Susanne E. Koster and David J. Langley[1]

TNO, Netherlands Organisation for Applied Scientific Research

Abstract

We investigate factors influencing choice behavior in online social networks. We use twitter data from a Dutch television talent show. In study one, we implement a nested conditional logit model with latent classes. We find heterogeneous effects. For two latent classes, cognitive factors most strongly influence choice behavior. For two other latent classes, the influence of the television show is strongest. For all latent classes, social effects positively influence individual choice. In study two, an agent-based model uses the parameters estimated in study one, and simulates marketing interventions. Our simulations show that collective choice behavior may be influenced by increasing the amount of time a product is displayed to the individuals, although changing the sequence that candidates appear or increasing the activity of influentials has no effect.

Keywords: Nested conditional logit, Latent classes, Agent-Based Model, Choice behavior, Online social networks, Opinion formation

[1] Corresponding author: david.langley@tno.nl

1 Introduction

Social networks and microblogs are now an important tool for information dissemination, sharing interpersonal communication and networking. The choices we make in our daily lives are increasingly influenced by these social networks. In addition to this, the amount, availability, and accessibility of social network data allows researchers to examine this field in its full complexity with ever increasing detail.

This research addresses the notion of individual choice behavior and the collective dynamics evolving from individual behavior in online social networks. We use twitter data related to a Dutch television talent show as a case study. The network of Twitter followers and followees represent the social network; the choices of the individuals to tweet about one of the candidates of the talent show represents the individual choice behavior, reflecting their opinion. We are interested in the factors that influence individual choice behavior in a social network. We review literature on social networks (Watts and Dodds (2007), Katona et al. (2011), Cha et al. (2010), Weng et al. (2010)), on new product diffusion (Bass (1969), Langley et al. (2012), Goldenberg et al. (2010, 2001)) and on social influence (Cialdini (2001), Leibenstein (1950), McPherson et al. (2001), Rafaat et al. (2009)). We distinguish between social network (the choice behavior of others in the social network), cognitive (the individuals' own characteristics), and external factors (influence of the television show). The first research question is:

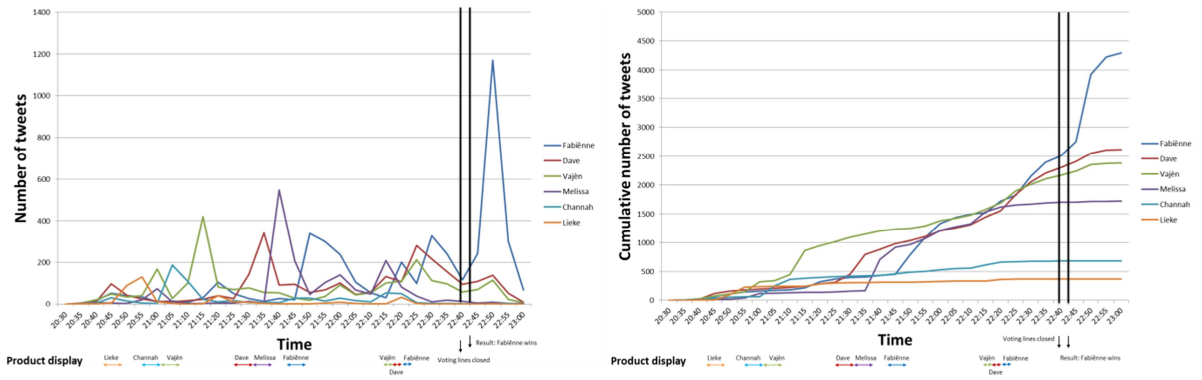
> *To what extent do social network, cognitive and external factors influence individual choice behavior in an online social network?*

Second, we are interested in how firms may intervene to change collective choice behavior in the online social network. For our case study this means that we alter parameters to make another candidate in the talent show more popular. We vary external factors (i.e. changing the sequences and frequency of the performances of the candidates), and we alter the choice behavior of influentials (i.e. making influentials with many followers tweet about a certain candidate). We believe that managers may influence these variables. The second research question is:

> *How may collective choice behavior be influenced by varying external factors and by altering the choice behavior of influentials?*

2 Data

Twitter data is collected on the Dutch television talent show The Voice Kids from the period February to March 2012. In this television show, the public votes on the best singing child. During the show, viewers are encouraged to tweet using the show's hashtag (i.e. #thevoicekids). We use Twitter API to stream 93,404 tweets about the program, sent by 20,822 individuals, who are connected by a network with 102,638 connections. Figure 1 shows the number of tweets per candidate (a) and the cumulative number of tweets per candidate (b) during the final show of The Voice Kids. Below the graph, the construct Product Display is shown, indicating when each candidate performs. The black vertical lines in the graphs indicate the time at which the voting was closed and the time at which the voting results were announced.



(a) (b)

Figure 1: Collective choice behavior: the number (a) and the cumulative number (b) of tweets per candidate during the final of the talent show The Voice Kids

The Twitter dataset consists of a list of tweets for which the content, the time stamp, and the sender is known. The followers of the sender are also known. This makes it possible to reconstruct the social network and to see in which time period individuals receive a tweet and when they send a tweet themselves. We are also able to see the content of the tweets. Based on the available data, we operationalize the constructs, as shown in Table 1.

Construct	**Operationalization**
Local Social Network effect	Number of tweets about a certain candidate or general tweets an individual receives from other individuals whom he or she follows, during the last hour
Global Social Network effect	Percentage of candidate tweets about a certain candidate during the last hour
Consistency	Total number of tweets about a certain candidate a user has sent in the past until now
Product display	1/0 variable indicating whether a candidates performs at time t

Table 1 Operationalization of the constructs

3 Study one: drivers of choice behavior

We implement a nested conditional logit model with latent classes. This model results in estimated parameters that indicate the relationship between individuals' choice behavior and its determinants. The parameters are latent class specific to allow for unobserved heterogeneity among individuals. The nested conditional logit model naturally clusters interrelated choices into nests. We distinguish the nest of sending tweets about candidates or sending a general tweet and the nest of sending no tweet. The probability to tweet about one of the candidates is defined as:

$$PR\left[Y_{it} = j \,\middle|\, x_{ijt}, C_{it} = 1\right] = \Pr\left[U_{ijt} > U_{ilt}\ for\ all\ l \neq j\right] = \frac{\exp(\alpha_{i,j} + x'_{ijt}\beta_i)}{\sum_{l=1}^{J} \exp(\alpha_{i,l} + x'_{ilt}\beta_i)}. \quad (1)$$

$for\ j = 1 \ldots J$ (for all candidates and the General Tweet option)
where,

- x'_{ijt} is a (1 x k) vector containing all k standardized explanatory variables described in Table 1.
- $\alpha_{i,j}$ is the individual and candidate specific intercept parameter.
- β_i is a (k x 1) vector of k coefficients.
- $Y_{it} = j\ for\ j = 1..J,$ means that user i tweets about candidate j at time t. If $Y_{it} = J$, a user sends a general tweet at time t.
- $C_{it} = m\ for\ m = 0, 1$. When $m = 1$, user i tweets about the talent show on time t. When $m = 0$, user i does not tweet about the talent show on time t.

The probability to send a tweet and to send no tweet is defined as:

$$PR[C_{it} = 1] = \frac{\exp(\tau_i I_{it})}{\exp(z'_{it}\gamma_i) + \exp(\tau_i I_{it})} \quad (2)$$
$$PR[C_{it} = 0] = \frac{\exp(z'_{it}\gamma_i)}{\exp(z'_{it}\gamma_i) + \exp(\tau_i I_{it})}$$

where,

- $I_{it} = \log\left(\sum_{j=1}^{J} \exp(\alpha_{i,j} + x'_{ijt}\beta_i)\right)$ is the inclusive value.
- z'_{it} is a (1 x r) vector containing all r standardized explanatory variables. The variables are the logarithm of the number of friends, and the logarithm of the number of followers. It is assumed that an individual is more active on Twitter, when this person has many friends and followers
- γ_i is a (r x 1) vector of r coefficients.

We implement latent classes to limit the number of parameters while capturing heterogeneity. We assume that there are S latent classes in the population. The overall probability that an individual belongs to class s equals $p_s = PR[s_i = s]$ with $\sum_{s=1}^{S} p_s = 1$. The individual-specific intercepts $\alpha_{i,j}$ will be replaced by a latent class specific intercept $\alpha_{s,j}$, the parameters β_i by β_s, γ_i by γ_s, and τ_i by τ_s. Let $\theta_s = \left(\alpha_{s,j}, \beta_s, \gamma_s, \tau_s\right)'$ be the set of parameters for class s. The likelihood contribution of individual i belonging to latent class s is given by (adapted from Franses and Paap (2010), pp. 108):

$$f\left(Y_{it} \mid x_{ijt}, z_{it}; \theta_s\right) = \prod_{t=1}^{T} \prod_{m=0}^{1} \prod_{j=1}^{J} \Pr[Y_{it} = j, C_{it} = m; \theta_s]^{I[C_{it}=m, Y_{it}=j\,]} \quad (3)$$

Write $\theta = (\theta_1 \ldots \theta_S)'$ and $p = (p_1 \ldots p_S)'$, the vector of the parameters for all S latent classes combined. Now, the likelihood function of the nested conditional logit model is given by:

$$L\left(\theta, p; Y_{it} \mid x_{ijt}, z_{it}\right) = \prod_{i=1}^{N} \left(\sum_{s=1}^{S} p_s \cdot f\left(Y_{it} \mid C_{it}, x_{ijt}, z_{it}; \theta_s\right) \right). \quad (4)$$

We use the EM algorithm as suggested by Wedel and DeSarbo (1995) to obtain parameter estimates.

3.1 Results

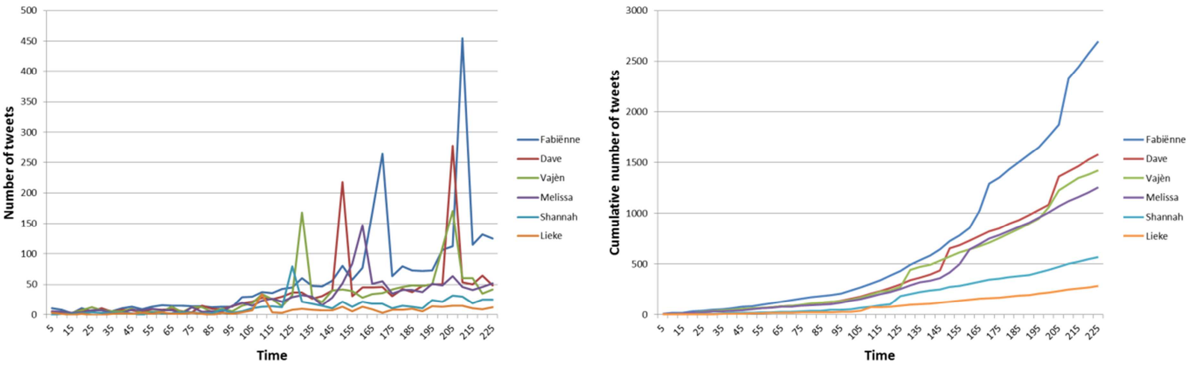
The minimum BIC value of The Voice Kids case study is obtained when 4 latent classes are implemented (McFadden R^2 is 0.247). Values from 0.2 to 0.4 are considered as highly satisfactory (McFadden, 1974, pp.121), and we therefore consider the model fit as highly satisfactory. Table 2 shows the parameter estimates for these four latent classes. We find support that social network, and cognitive factors positively influence individual choice behavior in a social network. The cognitive factor Consistency, i.e. the tendency to be consistent with previous choices, explains most of the variance in two out of four latent classes. The effect of the Television show explains most of the variance in the other two out of four latent classes. The Global Social Network effect, i.e. the effect of public opinion of the complete social network, explains third most of the variance. The Local Social Network effect, i.e. the effect of the choice behavior of direct connections, explains a smaller, but significant part of the variance in individual choice behavior. For the influence of external factors, i.e. moments at which the product is displayed to the individuals, we find mixed results.

Latent Classes	**1**	**2**	**3**	**4**
Local Social Network effect	0.001	0.028	0.099	0.249
Global Social Network effect	0.105	0.156	0.176	0.288
Consistency	0.133	0.538	0.221	0.942
Television show	0.191	0.091	0.235	0.336
Product Display	n.s.	-0.017	0.221	0.109
Log Number of Friends	-0.348	-0.236	-0.160	n.s.
Log Number of Followers	n.s.	0.252	0.243	0.801
Tau	7.667	7.438	4.921	6.512
p_s	0.490	0.283	0.181	0.045

Table 2 Parameter estimates of the Voice Kids case study per latent class, n.s. indicates not significant coefficients.

4 Study two: simulating effects of marketing interventions

We implement an agent-based model to simulate interacting heterogeneous agents in a social network. To build a rigorous agent-based model, we use the guidelines proposed by Rand and Rust (2011). The agent characteristics are drawn from the results as specified by the nested conditional logit model with latent classes. The estimated parameters are transformed into parameters that match unstandardized dependent variables. The model is programmed in Repast. During the simulations we notice the occurrence of unrealistic cascades, i.e. the number of one or more of the tweet options rapidly increases in time to a total number that is approximately 50-100 times larger than the average level of the total number of tweets. The cascades may be the consequence of non-stationary components in the nested conditional logit model. We decide to build in a cooling down period of 6 minutes, which ensures that an agent is not able to send a tweet again immediately after his previous tweet. Next, we compare the simulation output in terms of the simulated total number of tweets per choice option over time (Figure 2) to the empirical total number of tweets per choice option over time (Figure 1). In Figure 2(a) and (b) we see the number of tweets over time of The Voice Kids simulation with a cooling down period of 6 minutes. In The Voice Kids case the television start at time step 100. The peaks in Figure 2(a) show the moments when one of the candidates performs. Figure 2(b) shows that the ranking of the candidate in terms of the number of tweets is the same as in the empirical data (compare with Figure 1b). Also, the total number of tweets per candidate closely approximates the empirical number of tweets.



(a) (b)

Figure 2: Agent Based model simulation results: the number (a) and the cumulative number (b) of tweets per candidate for the talent show The Voice Kids with cooling down period of 6 minutes

We simulate managerial interventions as scenarios for The Voice Kids case. Table 3 summarizes the findings of the what-if scenarios for the agent-based model for The Voice Kids case. The simulations show that increasing the number of performances of the candidate (Product Display) influences the ranking of the candidates. The other two influence measures have no effect.

Influence measure	Result
Product Display frequency	Ranking of the candidates in terms of the number of tweets changes
Product Display sequence	No changes in the ranking of candidates
Activity of Influentials	No effect, Agent Based model becomes unstable

Table 3 The effect of influence measures for the Agent Based Model for The Voice Kids case

5 Discussion and managerial implications

The nested conditional logit model shows the underlying determinants of individual choice behavior. Managers may use the model for other applications, where individuals choose among a fixed set of products or candidates, for instance candidates in other television shows or in political elections. The latent classes of the nested conditional logit model take into account unobserved heterogeneity. For marketing managers a model with latent classes is useful in order to classify consumers or voters based on their actual behavior. Marketing managers may use this information for market segmentation and for the design of effective segment specific marketing campaigns. We show how managers may use the parameters of the nested conditional logit model as input for an agent-based model. Such a simulation model has empirically determined input parameters, and it has therefore a high input validity. Managers are able to test scenarios in a risk-free way, and still have trust in the validity of the results.

6 References

Bass, F. M., 1969, A new product growth for model consumer durables. *Management Science*, 15, pp. 215-227

Cameron, C., Trivedi, P.K., 2005, *Microeconometrics: methods and applications*, New York, Cambridge University Press

Cialdini, R. B. 2001. *Influence: science and practice*, Boston, Allyn & Bacon

Cha, M., Haddadi, H., Benevenuto, F. Gummadi, K.P., 2010, Measuring user influence in Twitter: The million follower fallacy, *ICWSM '10: Proceedings of international AAAI Conference on Weblogs and Social Media*

Franses, P.H., Paap, R., 2010, *Quantitative models in marketing research*, New York, Cambridge University Press

Goldenberg, J., Libai, B., Muller, E., 2001, Talk of the network: a complex systems look at the underlying process of word-of-mouth, *Marketing Letters*, 12,3, pp. 211-223

Goldenberg, J., Libai, B., Muller,. E. 2010, The chilling effects of network externalities, *International Journal of Research in Marketing*, 27, pp. 4-15

Heij, C., de Boer, P., Franses, P.H., Kloek, T., van Dijk, H.K., 2004, *Econometric methods with applications in business and economics*, New York, Oxford University Press

Katona, Z., Zubcsek, P., Sarvary, M., 2011, Network effects and personal influences: Diffusion of an online social network, *Journal of Marketing Research*, 48, 3, pp. 425-554

Langley, D.J., Bijmolt, T.H.A., Ortt, J.R., Pals, N., 2012, Determinant of social contagion during new product adoption, *Journal of Product Innovation Management*, 29, 4, pp. 623–638

Leibenstein, H., 1950. Bandwagon, Snob, and Veblen Effects in the Theory of Conspicuous Demand. *Quarterly Journal of Economics*, 64, 183-207

McFadden, D., 1979, Quantitative methods for analysis travel behaviour of individuals: some recent developments, In: Hensher, D.A. and Stopher, P.R.: (eds): *Behavioural travel modelling*, London: Croom Helm, pp. 279-318

McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a feather: Homophily in social networks, *Annual Review of Sociology*, 27(1), pp. 415-444

Rafaat, R.M., Chater, N., Frith, C., 2009, Herding in humans, *Trends in Cognitive Sciences*, 13 (10), pp. 420-428

Rand W., Rust, R.T., 2011, Agent-based modelling in marketing: guidelines for rigor, *International Journal of Research Marketing*, 28 (3), pp. 181-193

Scott, A.J., Wild, C.J., 1997, Fitting regression models to case-control data by maximum likelihood, *Biometrika*, 84, 1, pp. 57-71

Watts D.J., Dodds, P.S., 2007, Influentials, networks, and public opinion formation, *Journal of Consumer Research*, Vol. 34, December 2007, pp.441-458

Wedel, M., DeSarbo, W.S., 1995, A mixture likelihood approach for generalized linear models, *Journal of Classification*, 12, pp. 21-55

Weng, L. Flammini, A., Menczer, F., 2012, Competition among memes in a world with limited attention, *Scientific Reports*, 2, 335; DOI:10.1038/srep00335