Modeling User Behavior in Adoption and Diffusion of Twitter Clients

Conference Paper · October 2011

DOI: 10.1109/PASSAT/SocialCom.2011.95 · Source: DBLP

3 authors, including:

Elenna R. Dugundji
Centrum Wiskunde & Informatica

43 PUBLICATIONS **505** CITATIONS

SEE PROFILE

Michiel van Meeteren
Loughborough University

68 PUBLICATIONS **284** CITATIONS

SEE PROFILE

2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing

Modeling user behavior in adoption and diffusion of Twitter clients

Elenna R. Dugundji, Ate Poorthuis, and Michiel van Meeteren

***Abstract*—With the onset of ubiquitous social media technology, people leave numerous traces of their social behavior in – often publicly available – data sets. In this paper we look at a virtual community of independent ("Indie") software developers for the Macintosh and iPhone that use the social networking site Twitter. Using Twitter's API, we collect longitudinal data on network connections among the Indie developers and their friends and followers (approximately 15.000 nodes) and their Twitter behavior over a period of five weeks (more than 600.000 "tweets"). We use this dynamic data on the network and user behavior to analyze the adoption of Twitter client software.**

I. Introduction

WITHIN the Indie community, four prominent software developers have developed Twitter clients (Tweetie, Twitterrific, Twittelator, Birdfeed) that compete for adoption within the community. Apart from these Indie Twitter clients, members of the virtual community can choose from a range of clients that are developed outside of the Indie community (for example, Tweetdeck, Twitterfon) as well as the standard Web interface provided by Twitter. Generally, social networks and social capital are considered to be important factors in explaining the adoption and diffusion of behavior. Previous qualitative ethnographic evidence for our case study supports this view. [1] Using discrete choice analysis applied to longitudinal panel data, we are able to quantitatively test for the relative importance of global cultural discourse, taste-maker influence and other contextual effects, node level behavioral characteristics, socio-centric network measures and ego-centric network measures, individual preferences and social network contagion, in users' decisions of what client software they choose to interface to Twitter.

Importantly, we furthermore demonstrate a method using readily available software to estimate the size of the error due to unobserved correlated effects in users' choices. This is critical to test for in any application of multinomial logistic regression where social influence variables and/or other network measures are used as explanatory variables, since their use poses a classic case of endogeneity. We show that even in a seemingly saturated model, the log likelihood of the model fit can increase significantly by accounting for unobserved correlated effects. Furthermore the estimated coefficients in the uncorrected model can be significantly biased beyond standard error margins. Failing to account for correlated effects can yield misleading market share predictions for users' preferences for Twitter clients.

The paper is organized as follows. First a brief review of literature is presented describing what the paper brings to an existing stream of behavioral modeling research. Next the context of the case study is described: the Indie Mac community, the role of social media in the community in general and the role of Twitter in particular, and finally the role of taste-makers in this on-line ecosystem. We then proceed to review features of the data and some descriptive statistics of the choice alternatives that will guide our modeling efforts. Together the understanding of the context of the case study and the insights from the available data, lead us to define nine sets of different kinds of social and individual explanatory variables to explore in our model, with different functional forms. Estimation results are summarized. Finally, directions for future research efforts are outlined.

II. Discrete Choice with Social Interactions

A. Multinomial Logit Model

Discrete choice analysis allows prediction based on computed individual choice probabilities for heterogeneous agents' evaluation of alternatives. In accordance with notation and convention in Ben-Akiva and Lerman [2], the multinomial logit model is specified as follows. Assume a sample of N decision-making entities indexed (1,...,n,...,N) each faced with a choice among J_n alternatives indexed (1,...,j,..., J_n) in subset C_n of some universal choice set C.

The choice alternatives are assumed to be mutually exclusive (a choice for one alternative excludes the simultaneous choice for another alternative, that is, an agent cannot choose two alternatives at the same moment in time) and collectively exhaustive within C_n (an agent must make a choice for one of the options in the agent's choice set). In general the composite choice set C_n will vary in size and content across agents: not all elemental alternatives in the universal choice set may be available to all agents. For simplicity in this paper however, we will assume that the choices are available to all agents.

Let $U_{in} = V_{in} + \varepsilon_{in}$ be the utility that a given decision-making entity n is presumed to associate with a particular

Manuscript received July 22, 2011.

E. R. Dugundji is with Universiteit van Amsterdam, P.O. Box 16697, 1001 RD Amsterdam, Netherlands (corresponding author, tel: +31 20 692 5813; e-mail: e.r.dugundji@gmail.com).

A. Poorthuis was with Universiteit van Amsterdam, Netherlands. He is now with the University of Kentucky, Lexington, KY 40506-0027 USA (email: atepoorthuis@gmail.com)

M. van Meeteren is with Universiteit van Amsterdam, Netherlands (email: michielvanmeeteren@gmail.com)

978-0-7695-4578-3/11 $26.00 © 2011 IEEE
DOI

alternative i in its choice set C_n, where V_{in} is the deterministic (to the modeler) or so-called "systematic" utility and ε_{in} is an error term. Then, under the assumption of independent and identically Gumbel distributed disturbances ε_{in}, the probability that the individual decision-making entity n chooses alternative i within the choice set C_n is given by:

$$P_{in} \equiv P_n(i \mid C_n) = \Pr\left(V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}, \forall j \in C_n\right)$$
$$= \Pr\left[V_{in} + \varepsilon_{in} \geq \max_{j \in C_n}\left(V_{jn} + \varepsilon_{jn}\right)\right] = \frac{e^{\mu V_{in}}}{\sum_{\forall j \in C_n} e^{\mu V_{jn}}}$$

where μ is a strictly positive scale parameter which is typically normalized to 1 in the multinomial logit model.

The systematic utility is commonly assumed to be defined by a linear-in-parameters function of observable characteristics $\mathbf{S}_n$ of the decision-making entity and observable attributes $\mathbf{z}_{in}$ of the choice alternative for a given decision-making entity:

$$V_{in} = h_i + V\left(\mathbf{S}_n, \mathbf{z}_{in}\right) = h_i + \boldsymbol{\gamma}_i{}'\mathbf{S}_n + \boldsymbol{\zeta}_i{}'\mathbf{z}_{in}$$

The term h_i is a so-called "alternative specific constant" (ASC), as good practice to explicitly account for any underlying bias for one alternative over another alternative. In other words, h_i reflects the mean of ε_{jn} - ε_{in}, that is, the difference in the utility of alternative i from that of j when all else is equal. Since it is the difference that is relevant, for a general multinomial case with J alternatives we can define a set of at most J – 1 alternative specific constants.

The terms $\boldsymbol{\gamma}_i = [\gamma_{i1}, \gamma_{i2}, \ldots]'$ and $\boldsymbol{\zeta}_i = [\zeta_{i1}, \zeta_{i2}, \ldots]'$ are vectors of unknown utility parameters respectively corresponding to the relevant observable agent characteristics $\mathbf{S}_n$, and observable agent-specific attributes $\mathbf{z}_{in}$ of the choice alternatives. In general the utility parameters may take alternative specific values, however when there is no variation of the agent characteristics $\mathbf{S}_n$ across the choice alternatives, we can define a set of at most J – 1 vectors of alternative specific coefficients for the case of the $\boldsymbol{\gamma}_i$.

B. *Social Interactions*

An outstanding challenge in discrete choice analysis is the treatment of the interdependence of various decision-makers' choices [3,4]. Brock and Durlauf [5] introduce social interactions in multinomial discrete choice models by allowing a given agent's choice for a particular alternative to be dependent on the overall share of decision makers who choose that alternative. If the coefficient on this interaction variable is close to zero and not important relative to other contributions to the utility, then the distribution of decision-makers' choices will not effectively change over time in relation to other decision-makers' choices. However, if the coefficient on this interaction variable is positive and dominant enough relative to other contributions to utility, there may arise a runaway situation over time as all decision-makers flock to one particularly attractive choice alternative. In short, the specification captures social feedback between decision-makers that can potentially be reinforcing over the course of time. In diverse literature this is referred to as a social multiplier, a cascade, a bandwagon effect, imitation, contagion, herd behavior, etc. [6]

We introduce a social feedback effect among agents by allowing the systematic utility V_{in} to be a linear-in-parameter β first-order function of the proportion x_{in} of a given decision-maker's reference entities who have made this choice. Our model differs from the Brock and Durlauf model in that we consider *non-global* interactions. Agents see different proportions, depending on who their particular reference entities are. Additionally, we also consider various socio-centric and ego-centric network measures and other explanatory variables as contributions to the utility.

C. *Endogeneity*

One econometric issue that arises in empirical estimation of social interactions in discrete choice models using standard multinomial logistic regression however, is that the error terms are assumed to be identically and independently distributed across decision-makers. It is not obvious that this is in fact a valid assumption when we are specifically considering interdependence between decision-makers' choices. We might reason that if there is a systematic dependence of each decision-maker's choice on an explanatory variable that captures the aggregate choices of other decision-makers who are in some way related to that decision-maker, then there might be an analogous dependence in the error structure. Otherwise said, the same unobserved effects might be likely to influence the choice made by a given decision-maker as well as the choices made by those in the decision-maker's reference group, which is a classic case of endogeneity. The results and coefficients of such a model are likely to be biased. To try to separate out effects, it is therefore first and foremost critically important to begin with an as well-specified model as possible, making use of relevant available explanatory variables. [7]

Dugundji and Walker [8] illustrate issues in the empirical estimation of a discrete choice model with network interdependencies using mixed generalized extreme value model structures with pseudo-panel data. Several modeling strategies are presented to highlight hypothesized interaction effects. In absence of true panel data on interaction between identifiable decision-makers, they use a priori beliefs about the social and spatial dimension of interactions to formulate the connectivity of the network and use socioeconomic data for each respondent as well as the geographic location of each respondent's residence to define aggregate interactions by grouping agents into geographic neighborhoods and into socioeconomic groups where the influence is assumed to be more likely. Technically, however, interactions between identifiable decision-makers may also be modeled using the approach described given the availability of suitable data.

In our empirical case study on adoption of Twitter clients, we do indeed have available data on which identifiable agents (Twitter users) plausibly influence other identifiable agents' choices, and furthermore we have longitudinal panel data observing repeated choices by agents over time. In this paper with such rich data, we continue this exploration of issues in the empirical estimation of discrete choice models with social interactions. Since our data is fairly large -more than 10,000 agents- we argue that the effect of unobserved correlated effects as perceived by any given agent is normally distributed, but is the same for that agent over the fairly short time period of the data collection. This simplified assumption allows us to specifically control for correlations in the error structure, through the use of mixed multinomial logit models with panel effects. [9]

D. *Capturing Unobserved Correlated Effects*

Suppose each agent n makes a sequence of choices at a number of points in time indexed $(1,\ldots,t,\ldots,T_n)$. For our case study, we will consider a general case where the number T_n of decision-making moments per agent varies across agents. We introduce an additive, normally-distributed agent-specific error term for each alternative i as follows:

$$U_{int} = V_{int} + \varepsilon_{int} + \sigma_i \xi_{in} \;;\; \xi_n \sim N(0,I)$$

Conditional on ξ_n , the probability that agent n makes a particular sequence of choices over time $(i_1,\ldots,i_{Tn})$ is given by the product of the probabilities for agent n making each individual choice i_t:

$$P_n(i_1,\ldots,i_{T_n} \mid \xi_n) = \prod_{\forall t \in T_n} \frac{e^{\mu\left(V_{int}+\sigma_i \xi_{in}\right)}}{\sum\limits_{\forall j \in C_n} e^{\mu\left(V_{jnt}+\sigma_j \xi_{jn}\right)}}$$

The unconditional user choice probability is the integral of this product over all values of ξ_n

$$P_n(i_1,\ldots,i_{T_n}) = \int_{\xi_n} \prod_{\forall t \in T_n} \frac{e^{\mu\left(V_{int}+\sigma_i \xi_{in}\right)}}{\sum\limits_{\forall j \in C_n} e^{\mu\left(V_{jnt}+\sigma_j \xi_{jn}\right)}} N(0,I)\, d\xi_n$$

E. *Econometric Estimation with Simulation*

The unconditional choice probability is approximated through simulation for any given value of ξ_{in} as follows:

1) Draw a vector of values of ξ_n from $N(0,I)$ for each alternative in the choice set C_n, and label this $\xi_n^{\,r}$ with the superscript $r = 1$ referring to the first draw
2) Calculate the conditional user choice probability for the particular sequence of choices made by agent n with this draw
3) Repeat steps 1 & 2 for R total number of draws and average the results

$$\hat{P}_n(i_1,\ldots,i_{T_n}) = \frac{1}{R} \sum\nolimits_{r=1}^{R} \prod_{\forall t \in T_n} \frac{e^{\mu\left(V_{int}+\sigma_i \xi_{in}{}^{r}\right)}}{\sum\limits_{\forall j \in C_n} e^{\mu\left(V_{jnt}+\sigma_j \xi_{jn}{}^{r}\right)}}$$

If the estimated coefficients σ_i can be shown to be statistically insignificant, we assume that the hypothesized endogeneity has negligible effect.

III. Case Study

"One of the fantastic things about Twitter clients is how easy it is for users to jump from one to another. Just type in a username and password and off you go. It's possible for anyone to write a Twitter client nowadays and have the opportunity to completely blow everyone else out of the water. It's very exciting. Very democratic. And it certainly seems like everyone… is trying to do just that. I'm just happy to be part of it, I know the developers of other clients and I can say definitively that competition is making all of us write better apps."

--Macworld interview with Tweetie developer Loren Brichter, 24 April 2009

A. *The Indie Mac Community*

The "Indie Mac" developer community refers to a group of independent software companies that develop software for Apple's Macintosh platform. The majority of them are one-person shops, except for the more successful ones who sometimes have a few employees (although more than ten is rare). These companies sell their software to worldwide markets over the Internet, circumventing the traditional costs of physical production and distribution, which require substantial capital investments. Despite the fact that these software companies could regard each other as competitors, there is a lively interaction between them. This is done primarily through online means since they rarely are physically co-located. In time, a specific Indie culture and habitus has developed among them that guides interaction and informal social hierarchy between Indie developers [10]. They could be considered a virtual community of practice [11] where the specific habitus and related tacit knowledge embedded within it guides the ideas about how Mac software should look, feel and function.

B. *The Role of Social Media*

Social media can be considered the infrastructural backbone of the Indie community. When analyzed functionally, the online behavior of Indie developers broadly performs three functions: identification and socialization; satisfactions of informational needs; and marketing [1]. The first two functions are discussed in detail elsewhere [10,12]. Here we will highlight the role of the online network of Indie developers in the marketing of their software as this is inherently related to Twitter client choice.

The Indie market can -because of its lack of physical production and distribution costs- be considered a "long tail" market [13]. The product is available to everyone with an Internet connection, but people need to find it and appreciate it in order to buy it. This means that online exposure is the crucial factor in marketing and that the prime determinant of economic success is derived from getting your software known past a certain "tipping point" [14]. Because the signal-to-noise ratio is very high on the Internet, there is a high potential added value of peer recommendation of software products. Indies use echo marketing as a form of peer review. If a new software title is released, other developers endorse it if they appreciate it - and often only if they jointly appreciate the software title and the developer who made it. These endorsements go through the online network, often reaching the specialized journalists of the Macintosh world. Thus, the size and structure of the online network and the inclination of other developers and intermediaries to echo the message influence the economic success of a developer to a strong degree [10].

C. Twitter

The idea behind Twitter is that you can post messages with a maximum of 140 characters on the Internet, which subsequently can be read by everybody who is "following" you. You only get the messages from those whom you follow. This allows you to simultaneously broadcast a message to a lot of people while being able to limit the amount of information that reaches you. Within the Indie Mac community, Twitter gained a critical mass of users relatively early and its use was omnipresent at the time of field work [1] and still is today. One of the reasons for this fast adoption is that a well-known Indie Mac developer, Craig Hockenberry of the Iconfactory, developed a desktop client for Twitter early on [15], which let Twitter run in the background so the user could concentrate on other things. After the advent and take off of the iPhone, a lot of Twitter activity has moved to that platform. A variety of Twitter clients is now available for the iPhone and there is a fast paced but friendly-voiced competition on innovation going on between them [16].

D. The Role of Taste-makers

Within such a business environment, it is evident that social capital plays a role in the economic chances of a company. Having your software endorsed by the community can be a great asset in economic terms. This endorsement usually follows from complying with the aesthetic and social discourses that guide "proper" behavior in the community. The literature on the cultural industries has emphasized the role of taste making actors in this respect [17]. A taste-maker could be defined as an intermediary actor who yields -often symbolic- power and uses that power, by for example endorsements, to help selected authors to become economically successful. Respondents acknowledged that these "taste-makers" play an important role in the Indie Mac community, especially because these persons often function as a bridge between the in-group of developers and the first tier of critical users.

One person who arose as an important taste-maker is technology journalist John Gruber, who maintains the Daring Fireball blog (http://daringfireball.net). Gruber, who has an educational background in computer science, is considered to be an important software "connoisseur" and tech industry insider by both Indie developers and the wider audience. During the data collection period he had more than 30,000 followers on Twitter while following less than 300 people, most of whom were reciprocal.

IV. Understanding the Data

"Perhaps the most important factor that has made Twitter such a rich category for client software is that there is so little friction to switch between apps. There's nothing to import or export, and zero commitment."

--John Gruber, Daring Fireball, 24 April 2009

A. Features of Twitter Data

This paper studies the adoption and diffusion of Twitter clients within the Indie community. Based on earlier research [18] we were able to determine a community of Indie developers that are actively using Twitter, using a mixed method community detection approach. For this community we use Twitter's publicly available API to gather data on network connections and actual messages sent. For 39 days, from 9 August until 16 September 2009, we harvested tweets and network connections on a daily basis for each of the nodes in the community. In this period the community sent a total of approximately 1 million messages of which 630633 are general posts.

Twitter networks are directed. A Twitter user can "follow" other users and have "followers." The distinction between the two is important as it influences how information on Twitter flows. This can best be explained with a practical example. Let's take two users: Joe User and John Gruber. Joe User is interested in Gruber and chooses to follow him. This means that as soon as Gruber sends a tweet, this tweet is also sent to Joe User. Joe User is thus continuously updated about everything that Gruber writes and is a "follower" of Gruber. However, following does not have to be reciprocal. If Gruber thinks that Joe User is not interesting, he does not have to follow back.

In short, the key aspects of this data are that we have:

--Large, longitudinal behavioral panel data: We can observe the repeated behavior of many users over (continuous) time.

--Explicit, directed networks: We know who is connected to whom, as well as the directionality of the flow of information.

B. Structuralist versus Connectionist Paradigm

For each tweet, we know the following information: User ID; Date and time sent; and the Twitter client that was used to send the tweet. In addition, we know which users are following the sending user and thus received the tweet. There are thus in principle over the duration of the observation period, two distinct types of "networks":

--An adjacency network of users which is a binary graph of 0's and 1's (a so-called "digraph") indicating for each user which other users they have told Twitter that they want to follow.

--A flow network of tweets which is a valued graph counting how many tweets are sent from one user to another in a particular time frame. For example tweets may be counted per day, or per week, or the total tweets may considered over the entire observation period.

The digraph corresponds to a so-called "Structuralist" or topology-based view of social network analysis; the flow network corresponds to a so-called "Connectivist" view. The digraph emphasizes ties between users as "roads" that simply exist or not, where "the structure of the ties in the network has profound effects of the capabilities and constraints of the network." [19] The flow network represents ties between users as the amount of "traffic" over the roads, where the emphasis lies in the transmission or diffusion of something across the ties, such as information, material, attitudes, behavior or resources.

In our case, since we are interested in modeling user behavior in the adoption and diffusion of Twitter clients, the Connectivist approach is most relevant. In our data preparation, we will use the knowledge of the adjacency network and the knowledge of who sent tweets on a given day to generate a series of daily tweet flow networks.

C. Tweeto Ergo Sum ("I Tweet Therefore I Am")

Important however, is that not all users in the adjacency network actually tweet during observation period. If a user does not tweet, the choice of twitter client cannot be observed. Consequences of this are:

--This user is effectively invisible to the researcher, since there is no observed behavior from this user.

--This user has no possibility to influence other users via tweet behavior.

Technically, this implies that the tweet flow matrices have less nodes (and less edges) than the adjacency matrix.

D. Rapid Adjustment and Negligible Transaction Costs

Every Twitter client emphasizes different functionality and has a slightly different design. In general, it can be argued that especially on mobile devices, there is a trade-off in User Interface (UI) design between complexity and usability. The larger a client's functionality the more complex -and thus less usable- its UI tends to become. Despite these differences it is important to stress that there are hardly any transaction costs associated with switching clients. Twitter messages are furthermore not stored within a client but remain on Twitter's servers. This means that there is no need for data migration when switching clients, and thus rapid adjustment. During the data collection period, we found that users in the Indie community used on average at least 3 different clients. [20]

E. Descriptive Statistics

Table I shows the frequency distribution of Twitter clients in the 630633 general post tweets during our data harvesting period from 9 August to 16 September 2009, and Sysomos data [21] on the distribution of Twitter clients for more than 500 million tweets for all Twitter users in the half-year period June-November 2009. If we compare client usage within our community with client usage in the larger public, we see the average use of Twitter's native web interface is dominant outside of the Indie community, whereas Indies tend to use a third party client more often. This might be explained by the fact that Indies are Twitter "power" users rather than recreational users, leading to the usage of more sophisticated technology in order to integrate Twitter in their daily workflow. Moreover, Tweetdeck is by far the most popular third party Twitter client in the entire population but Indies make heavy use of Tweetie, which is used almost three times as often as the web interface.

TABLE I
MARKET SHARE OF MOST-USED TWITTER CLIENTS

Sysomos Data		Indie Mac Community	
Twitter Client	Market Share	Twitter Client	Market Share
Web	46,8%	Tweetie	45,2%
Tweetdeck	8,5%	Web	12,8%
Tweetie	2,8%	Twitterrific	7,6%
Twitterrific	1,6%	Tweetdeck	6,4%
Seesmic	1,1%	Twitterfon	2,5%
Twitterberry	1,0%	Twittelator	1,5%
Hootsuite	0,6%	Birdfeed	1,1%
Other	37,5%	Other	22,9%

"Web" refers to the default Twitter web interface

V. MODELING THE EFFECTS

Based on our review of the case study and the data, we expect client choice to be influenced by a number of distinct dimensions. Generally, social networks or social capital are considered to be important variables in explaining the adoption and diffusion of behavior. However, it is debatable to what extent the actual social connections, the global cultural discourse, and individual preferences influence this adoption and diffusion. Through our modeling of the effects, we can try to test the different hypotheses.

We distinguish between four Indie clients (Tweetie, Twitterific, Birdfeed and Twittelator), two popular non-Indie clients (Twitterfon and Tweetdeck) and the default Twitter web interface ("Web"). In addition, we employ a choice alternative, "Other" that serves as a baseline reference for the modeling. The "Other" category is highly heterogeneous and consists of more than 3500 clients that have relatively small market share (< 1%). On the basis of data we proceed to construct the following nine sets of different kinds of social and individual explanatory variables to explore in our model.

A. Contextual Effects: Taste-Maker Influence

"Birdfeed… has some features which you can't believe aren't in every iPhone Twitter client."

--John Gruber, Daring Fireball, 29 June 2009

We start with exploring the contextual effect of whether or not a user in the community is connected to professional independent tech blogger John Gruber. Since Gruber promotes different clients to different extents [20], we are interested to see if the clients he promotes most favorably are used more often by the users connected to him. We operationalize this dummy variable in two different ways: if a user "follows" Gruber (ie. user receives tweets from Gruber); and if there is a reciprocal link with Gruber (ie. the link with Gruber is especially strong).

B. *Contextual Effects: Developer Influence*

Next, we are interested in the contextual effect of whether or not a user in the community is connected to a Twitter client developer [20] as follows: *Clients developed by "Indies*": Tweetie (Loren Brichter, Atebits); Twitterrific (Craig Hockenberry, Iconfactory); Twittelator (Andrew Stone, Stone Design); Birdfeed (Buzz Andersen, SciFi HiFi); *Clients developed by others:* TweetDeck (Iain Dodsworth, TweetDeck); TwitterFon (Kazuho Okui, Naan Studio). We operationalize each of these dummy variables in two different ways: if a user "follows" the developer (ie. user receives tweets from the developer); and if there is a reciprocal link with the developer (ie. the link with the developer is especially strong).

C. *Behavioral Characteristics: Power Users*

"Different people seek very different things from a Twitter client. TweetDeck, for example, is clearly about showing more at once. Tweetie is about showing less. … There is so much variety because various clients are trying to do very different things."

--John Gruber, Daring Fireball, 24 April 2009

Since the Twitter clients have very different features, we might expect users who tweet a lot to prefer different kinds of clients than users who tweet less frequently. We operationalize this variable in four different ways: number of tweets sent by a user during observation period; "status count" (total tweets sent by a user during their entire history); number of tweets sent by a user *prior* to observation period (ie. giving emphasis of how active the user was in the past and how long the user has been using Twitter); and finally, the ratio of tweets sent by a user during observation period to total tweets sent during their entire history.

D. *Network Measures: Central Users*

As per our review of the importance of social media networks for "echo-chamber" marketing, we are interested in whether a user's position in the community affects their client choice. We compute five classic network centrality measures [22, 23]: in-degree centrality (the number of a user's "friends" in sample, ie. from whom tweets are received); out-degree centrality (the number of a user's "followers" in sample, ie. to whom tweets are sent); closeness centrality (sum of distances from a user to all other users, giving an indication of the expected time until arrival for information that might be flowing through the network); betweenness centrality (how often a user lies along the shortest path between two other users, giving an indication of access to diversity of information); and finally, eigenvector centrality (measures if a user is connected to many users who are themselves well connected, identifying users in centers of cliques).

E. *Network Measures: Extended User In-Degree*

In order to test the relative importance of the exposure to information flowing through the wider Twitter universe outside of the Indie community, we explore three extra network measure variables: the total number of a user's "friends" in the entire Twitter universe, ie. from whom a given user in principle receives tweets; the number of users *outside* the community from whom a given user in principle receives tweets; and finally, the ratio of users inside sample from whom a given user receives tweets to their total "friends" in the Twitter universe.

F. *Network Measures: Extended User Out-Degree*

Similarly, in order to test the relative opportunity to influence other users in the wider Twitter universe outside of the Indie community, we explore three extra network measure variables: the total number of a user's "followers" in the entire Twitter universe, ie. to whom a given user in principle sends tweets; the number of users *outside* the community to whom a given user in principle sends tweets; and finally, the ratio of users inside sample to whom a given user sends tweets to their total "followers" in the Twitter universe.

G. *Temporal Effects: Individual Preferences*

"There are several factors that make Twitter a nearly ideal playground for UI design. The obvious ones are the growing popularity of the service itself and the relatively small scope of a Twitter client. Twitter is such a simple service overall, but look at a few screenshots of these apps, especially the recent ones, and you will see some very different UI designs, not only in terms of visual style but in terms of layout, structure, and flow. … It is not easy to write a good client for something as small in scope."

--John Gruber, Daring Fireball, 24 April 2009

We operationalize individual preference by constructing an alternative-specific relative individual cumulative lag variable. For each tweet, we count how often the sending user has been using each client in the 7 days prior to sending the tweet resulting in an absolute cumulative lag variable. For each client, we then convert this absolute frequency to a relative cumulative lag variable indicating that client's use relative to how often that user has been using other Twitter clients in the past 7 days. This individual preference variable shows how "sticky" a particular client has been for a user in the past 7 days. This individual past behavior is likely to be a predictor of client choice for the next tweet, capturing complex UI preferences which we as researchers were not able to measure directly.

H. Temporal Effects: Social Network Contagion

"Gruber loves Tweetie
So do all of the cool kids
Give me your money"

--Haiku by Tweetie developer Loren Brichter, in the tweet-sized advertisement to promote "Tweetie for Mac"

To operationalize network influence we use the absolute cumulative lag variable as a basis. For each tweet, we count how often all users that the sender of that specific tweet is following use each client in the 7 days prior to sending that the tweet. We convert the absolute frequency to an alternative specific relative network influence variable that indicates how often each client has been used relative to all other clients by all users that the sender of the tweet is following (ie. receiving information from). This can entail specific mentions of a client in a tweet but also more implicit or tacit knowledge about which client is popular or deemed useful within that user's social network. We argue that this usage by "friends" might influence client choice by either specific mentions of a client in Tweets or by the effect of tacit knowledge encoded within a user's social network.

I. Global influence

The cultural discourse of what is popular within the entire Indie community is operationalized by a set of alternative specific constants (ASC). Amongst things such as price and the impact of media exposure, we argue that this effectively captures global influence. Similar to an intercept in linear regression, it indicates the popularity of an alternative relative to all other alternatives during the entire sample period, after controlling for all other effects.

VI. Putting It All Together

All models are estimated using the freely available optimization toolkit Biogeme (http://biogeme.epfl.ch) developed by Bierlaire [24]. We begin by estimating a baseline multinomial logit model with alternative specific constants only, representing global bias. The log likelihood, number of estimated parameters and adjusted rho-squared are given in the first line of Table II. Since we have defined 8 choice alternatives, there can be at most 7 alternative specific constants (as explained in Section II.A).

Next we test one-by-one each of the explanatory variables defined in Section VI.A–H. In cases where the variables are continuous (ie. for all cases except for the dummy variables in section VI.A and VI.B), we also test linear, quadratic and square root forms of these variables. Based on log likelihood tests compared to the baseline model and t-tests on the estimated coefficients [2], we identify the best fitting variables per category. For example, the dummies defined as "follows Gruber" and "follows developer" are more significant than their respective forms "reciprocal link with Gruber" and "reciprocal link with developer"; the most significant centrality measures are closeness and square root of eigenvector centrality, etc. The interested reader is referred to [20] for details and interpretation.

Having determined the best fitting variables and their respective functional forms, we then add the variables incrementally to the model, testing the improvement in log likelihood at each step. This is important to do, since variables that may have been significant when included in the model specification on their own, might no longer be significant when included together due to significance being shared between variables. The results are reported in lines 2-10 of Table II. Each successive specification adds 7 new parameters to the model (with the exception of "follows developer" where there are 6 since the Web alternative does not have a third party developer), as our data is rich and extensive enough to support alternative-specific definitions of the variables. In our case study, each new set of variables significantly improves the log likelihood (p-value of 0.000).

TABLE II
Log Likelihood Tests for Incremental Model Specifications

Nr.	Log Likelihood	Est. Par.	Rho Sq.	-2* $[L_R-L_U]$	χ2 (0.01)	p-Value
1	-968350,6	7	0.262	- -	- -	- -
2	-954368,7	14	0.272	27964	18.5	0.000
3	-568721,3	21	0.566	771295	18.5	0.000
4	-567945,3	28	0.567	1552	18.5	0.000
5	-566798,7	34	0.568	2293	16.8	0.000
6	-562010,9	41	0.571	9576	18.5	0.000
7	-561154,7	48	0.572	1712	18.5	0.000
8	-560664,6	55	0.572	980	18.5	0.000
9	-559662,6	62	0.573	2004	18.5	0.000
10	-559546,0	69	0.573	233	18.5	0.000
11	-452048,1	76	0.655	214996	18.5	0.000

1: Baseline model with alternative-specific constants only; 2 + Social network contagion (sq root); 3: + Lagged individual preferences (sq root); 4: + Follows Gruber; 5: + Follows developer; 6: + Frequency tweets during observation period (sq root); 7: + Eigenvector centrality (sq root); 8: + Closeness centrality; 9: + Ratio in-degree to total friends in Twitter (sq root); 10: + Ratio out-degree to total followers in Twitter (sq root); 11: + Estimated user-specific error component

Finally, we include the normally-distributed user-specific error terms as in Section II.D. We test the robustness of results using three different optimization algorithms for the maximization of the log likelihood, each with 10 different random seeds for generating the draws. We use the estimated coefficients from the model in line 10 of Table II as a starting point for these 30 estimation runs with 50 draws, and then use the results with 50 draws in turn as the starting point for another 30 estimation runs with 200 draws, etc., for increasing number of draws, until the results stabilize across the random seeds for the three different optimization algorithms. Accounting for the unobserved correlated effects gave a dramatic jump in log likelihood as seen in line 11 of Table II. The estimated coefficients in the final model in line 11 were also significantly different beyond standard error margins for 63 of 69 variables in the model in line 10. [20] Failing to account for unobserved correlated effects can thus yield misleading market share predictions for users' preferences for Twitter clients.

VII. Conclusion and Recommendations

A prominent approach to studying the dynamics of networks and behavior stems from a growing stream of research on stochastic actor-based models. See Snijders, van de Bunt, and Steglich [25] for a thorough tutorial. With the large data in our case study however, these established methods are not tractable. The alternative approach we discuss in this paper allows us to apply other freely available, open source, existing software for the estimation of the models. In so doing, we hope to stimulate researchers and practitioners to adopt these techniques when using large data sets of more than 1000 nodes due to the relatively lower entry barrier than could be the case if dedicated code would need to be written or if expensive software would need to be purchased.

Our focus can be placed into broader context by drawing attention to three hypotheses on social interactions highlighted by Manski [6] to explain the common observation that individuals belonging to the same group tend to behave similarly: endogenous effects, contextual effects, and correlated effects. The first two hypotheses express inter-agent causality in a model; the third hypothesis does not. A key distinction between the two inter-agent causal effects is that the first involves feedback that can be reinforcing over the course of time. Aral, Muchnik and Sundararajan [26] adapt propensity score matched sample estimation [27] for use in dynamic networked settings, providing an empirical application to the adoption of a mobile service application for the global instant messaging network Yahoo!. An interesting direction for further discrete choice research on adoption and diffusion in large networks may be combining the approach of Aral *et al* for distinguishing causal effects, with the present work accounting for unobserved correlated effects.

Acknowledgment

Earlier versions of this on-going research were presented at: 10th International Conference on Computational Science (ICCS), Universiteit van Amsterdam, Netherlands, 31 May-2 June 2010; XXX "Sunbelt" Conference of the International Network for Social Network Analysis, Riva del Garda, Italy, 29 June-4 July 2010; XVII World Congress of Sociology of the International Sociological Association (ISA), Gothenburg, Sweden, 11-17 July 2010; 2nd Workshop on Information in Networks (WIN), New York University, USA, 24-25 September 2010; 4th International Conference on Computational and Financial Econometrics (CFE), University of London, UK, 10-12 December 2010; and the Interdisciplinary Workshop on Information and Decision in Social Networks (WIDS), Massachusetts Institute of Technology, USA, 31 May-1 June 2011.

References

[1] M. van Meeteren. (2008, July 14). Indie fever: The genesis, culture and economy of a community of independent software developers on the Macintosh OS X platform. *A Sofa Publication* [On-line]. Available: http://www.madebysofa.com/indiefever

[2] M. Ben-Akiva and S. R. Lerman, *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press, 1985.

[3] D. McFadden, "Economic choices," *American Economic Review*, 91(3), pp. 351-378, 2001.

[4] D. McFadden, "Sociality, rationality, and the ecology of choice," in *Choice Modelling: The State-of-the-Art and the State-of-Practice*, S. Hess and A. Daly, Eds. Bingley, UK: Emerald Group Publ. Ltd, 2010.

[5] W. A. Brock and S. N. Durlauf, "A multinomial choice model of neighborhood effects," *American Economic Review*, 92(2), pp. 298-303, 2002.

[6] C. F. Manski, *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard Univ. Press, 1995.

[7] E. R. Dugundji, "Socio-Dynamic Discrete Choice," Ph.D. manuscript. Dept. Geography, Planning and International Development Studies, Universiteit van Amsterdam, Netherlands, 2011.

[8] E. R. Dugundji and J. L. Walker, "Discrete choice with social and spatial network interdependencies: An empirical example using mixed generalized extreme value models with field and panel effects," *Transportation Research Record*, 1921, pp. 70-78, 2005.

[9] K. E. Train, *Discrete Choice Methods with Simulation*, 2nd ed. New York: Cambridge University Press, 2009.

[10] M. van Meeteren, "An ethnography of Indie software developers," unpublished.

[11] A. Amin and J. Roberts, "Knowing in action: Beyond communities of practice," *Research Policy*, 37(2), pp. 353-369, 2008.

[12] J. S. Brown and P. Deguid, *The Social Life of Information*. Boston: Harvard Business School Publishing, 2000.

[13] C. Anderson, *The Long Tail*. London: Random House Books, 2006.

[14] M. Gladwell, *The Tipping Point*. London: Abacus, 2000.

[15] The Iconfactory. (2011). Twitterrific version history. [Online]. Available: http://iconfactory.com/software/twitterrific_history

[16] K. Baxter. (2009, May 7). *Tightwind* [Online]. Available: http://www.tightwind.net/2009/05/another-example-of-the-greatness-that-is-the-mac-community

[17] R. E. Caves, *Creative Industries*. Cambridge, MA: Harvard Univ. Press, 2000.

[18] M. Meeteren, A. Poorthuis, and E. R. Dugundji, "Mapping communities in large virtual social networks," presented at Engaging Data: First International Forum on the Application and Management of Personal Electronic Information, MIT, Cambridge, MA, Oct. 12-13, 2009.

[19] S. P. Borgatti, "Social network analysis: Overview of the field today," keynote address presented at National Academy of Sciences, Sept. 2005.

[20] E. R. Dugundji, A. Poorthuis, and M. van Meeteren, "Capturing unobserved correlated effects in diffusion in large virtual networks: Distinguishing individual preferences, social connections and cultural discourse influence on the adoption of Twitter clients," unpublished.

[21] Sysomos Inc. (2009, Nov.). Inside Twitter clients. [Online]. Available: http://www.sysomos.com/insidetwitter/clients

[22] L. C. Freeman, "Centrality in social networks," *Social Networks*, 1(3), pp. 215-239, 1978/79.

[23] P. Bonacich, "Power and centrality: A family of measures," *Am. J. Soc*., 92(5), pp. 1170-1182, 1987.

[24] M. Bierlaire, "Biogeme: A free package for the estimation of discrete choice models," presented at the 3rd Swiss Transportation Research Conference, Ascona, Switzerland, 2003.

[25] T. A. B. Snijders, G. G. van de Bunt, and C. E. G. Steglich, "Introduction to stochastic actor-based models for network dynamics," *Social Networks*, 32(1), pp. 44-60, 2010.

[26] S. Aral, L. Muchnik and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proc. Nat'l Acad. Sci. USA*, 106(51), pp. 21544-21549, 2009.

[27] S. Hill, F. Provost, and C. Volinsky, "Network-based marketing: Identifying the likely adopters via consumer networks," *Statistical Science*, 21, pp. 256-276, 2006.