Eric M. Schwartz
ericschw@wharton.upenn.edu
February 27, 2009

1st Year PhD Student, Marketing
Co-Advisors: Eric Bradlow and Peter Fader
2009 Ackoff Doctoral Student Fellowship

How Much Social Network Data Do We Really Need?

Project Summary

What data should firms use to extract value from their massive (and growing) databases of individual customer transactional histories, online social connections, demographics, and marketing activity to make their decisions in social media? Researchers and firms have gotten too caught up analyzing the nature of the social *networks* as an end onto itself, rather than focusing on behavioral outcomes such as application downloads, site visits, etc.. Is the full network information necessary to answer the key questions that inform a firm's decision process? Are data about every customer connection really necessary to forecast revenue-driving behavior and to strategically target the most valuable and "influential" customers?

As a first step, it is critical understand the fundamental *individual-level* behaviors that make up social media and commerce. Parsimonious stochastic models of such core behavioral components, such as timing (waiting to adopt, waiting to dropout), counting (using a service, sharing it with others), and choice (accepting messages) processes traditionally explain a great deal of variation in customer activity in countless domains in marketing. Why should we expect them to perform differently in the domain of online social media? By beginning with a parsimonious benchmark model and minimal data, we can rigorously test (in a nested fashion) the incremental importance of adding more complicated model structure (via network data) to answer scientifically interesting questions of consumer behavior that will directly impact firm decision making.

Under the supervision of Professors Eric Bradlow and Pete Fader and with the collaboration of corporate participants in the Wharton Interactive Media Initiative (WIMI), I plan to do the following: (1) integrate fundamental stochastic models to build a probability model of individual behavior; (2) establish minimal data and sufficient statistics to answer key managerial questions; (3) provide out-of-sample forecasts of future revenue-generating activity; (4) run field experiments with a company managing various social media services including Facebook applications with millions of unique users; and later, (5) incorporate a utility-based approach to explore how why consumers actively and explicitly influence others in an online setting through inviting, sharing, gifting, referring, and recommending. I hope to use the fundamental probability modeling approach to build a baseline model upon which I will build richer stories with economic and psychological motivations.

Key managerial questions guiding firms' decision processes and risk management still remain unresolved. These are drawn from the traditional customer base analysis of customer lifetime value.

- Which consumers will be most valuable to the firm next period? Should the firm target them?
- Which consumers should we not bother targeting (either because they are going to be active anyways or because they have already "dropped out")?
- Given consumers' past social media behavior, how many times can we expect consumers to use the product next period? How many new users can we expect will each consumer to bring in?
- Given their past behavior facing the firm's marketing activity, how many times can we expect consumers to use the product facing future marketing activity next period?

A social media company managing various Facebook applications has already provided us with their full database, and we have begun planning field experiments. While many researchers have called for the need of such data (Hill et al. 2006), to our knowledge, the present research will be the first to use social media data with all of the following ingredients:

- Individual-level longitudinal data of explicit one-to-one communication between customers exclusively about the product of interest (i.e., invitations, referrals);
- Revenue generated directly from each individual's behavior;

- Complete network data from product launch for two years (i.e., no sampling, no left censoring);
- Large-scale controlled field experiment (i.e., not just field study or natural experiment).

With great data comes great responsibility. Instead of running a saturated model, full of all "bells and whistles" used in social network analysis, we start from a clean slate. This complete data gives us the flexibility to "squash" the data to find the least possible data needed to answer each question. That is, we will take marginals of the matrix to find the sufficient statistics.

The goal is to minimize data and maximize explanatory power. How much are we minimizing the data? The minimal data are simply a few pieces of information for each individual. In our data for one Facebook application, the minimal data are: (1) how long did it take to adopt after receiving a first invitation, (2) frequency of use, (3) recency of application use. The "full data matrix" typically used to store information on social interactions is a three-dimensional structure where the (i,j,t)-th entry indicates the direction (and strength) of social interaction between any two users in a short time. If the time is represented by 52 discrete weeks and 1,000,000 unique users have adopted the service, then there are 10^6 x 10^6 x 52 data points. Since time is not really in discrete buckets, the true 10^6 x 10^6 interaction matrix evolves in continuous time. The data are unwieldy to say the least. Clever attempts have been made to reduce the computational burden of looking at the formation of all of these links, but those attempts still hold the dyadic links as sacrosanct (i.e., Braun and Bonfrer 2009; Ansari and Koenigsberg 2009). We minimize data using the individual as the core unit of analysis.

Recent work has been dedicated to econometric concerns of identification of social influence, including endogenous group formation, correlated unobservables, and simultaneity (Hartman et al. 2005; Nair et al. 2008). Many of these concerns (but not all) go away when one has the "right" data. We echo the work of researchers carefully attending to the econometric concerns of identification. Just because two consumers are in the "same social network" or have communicated previously does not mean there is causal social contagion between them. True "contagion through word of mouth" however, can be inferred from the sequential nature of the present data: one consumer actively sends an invitation for a service to another consumer who has not yet received an invitation from anyone else. The recipient accepts the invitation, adopts the service, and begins generating revenue for the firm.

I intend to use this project to build a foundation for a stream of cumulative research. Future models should pass the test of contributing knowledge above and beyond this parsimonious customer base analysis model composed of simple stochastic building blocks (count, timing, choice processes). Such future models will allow for richer stories with economic and psychological motivations on both the firm and consumer side. Still using the minimal data, a latent variable model will allow us to examine a correlation structure linking the various processes (i.e., people who adopt quicker, use the product more, but drop out sooner). Expanding beyond the minimal data, we can specify a process for consumers bringing in new users. Even further, our benchmark model will serve as a higher bar for testing behavioral theories. We establish the "null hypothesis" for future tests. In other words, this approach is a more sophisticated model of the "intercept" of future regressions. Here, demographics and network characteristics of individuals can be incorporated. From here, we can further tackle the concern that consumers "closer" in a latent network space are more likely to interact than consumers "farther apart." In addition, we can specify a utility function and "social effort" constraint of time driving a consumer's decision process for adopting, using, sharing, and ending activity. While it is possible such an economic story may not improve forecasts of the baseline model, it will provide scientifically interesting and relevant insight that the earlier attempt does not offer.