Mark Lett (2011) 22:15–29
DOI 10.1007/s11002-009-9099-9

A social network analysis of customer-level revenue distribution

Michael Haenlein

Published online: 12 January 2010

Abstract Social network analysis has been a topic of regular interest in the marketing discipline. Previous studies have largely focused on similarities in product/brand choice decisions within the same social network, often in the context of product innovation adoption. Not much is known, however, about the importance of social network effects once customers have been acquired. Using the customer base of a telecommunications company, our study analyzes network autocorrelation in the distribution of customer-level revenue within a social network. Our results indicate a significant and substantial degree of positive network autocorrelation in customer-level revenue. High (low) revenue customers therefore tend to be primarily related to other high (low) revenue clients. Furthermore, we show that approximating communicative proximity by spatial proximity leads to a substantial underestimation of these effects.

Keywords Social network analysis · Spatial statistics · Network autocorrelation · Moran's *I* · Gary's *C*

1 Introduction

Interdependencies in consumer behavior among people who are connected to each other, or, more generally, the topic of social network analysis, have been of steady interest in marketing literature for over 25 years. Previous research has largely focused on brand choice decisions and brand congruence in interpersonal relations (Reingen et al. 1984) or, more generally, on product choice, specifically in the context of innovation diffusion (Goldenberg et al. 2009). These studies have provided consistent support for the fact that people who are related to each other by

The author thanks Roger Bivand, Professor of Economics at the Norwegian School of Economics and Business Administration and developer of the spdep package, for his helpful comments during data analysis.

M. Haenlein (✉)
ESCP Europe, 79 Avenue de la République, 75011 Paris, France
e-mail: haenlein@escpeurope.eu

strong tie relationships tend to show similarities in consumption behavior and brand preferences, caused by factors such as word-of-mouth communication (Goldenberg et al. 2001), social influences (i.e., compliance, identification, and internalization; see Kelman 1961), market embeddedness (i.e., utility derived from social capital next to basic product attributes; Frenzen and Davis 1990), or the use of similar information sources (Goldenberg et al. 2009). This finding has direct implications for the customer acquisition process and can, for example, be used in the context of network-based marketing campaigns (Hill et al. 2006).

Customer acquisition or relationship initiation is, however, only one part of the firm's CRM process, which also includes relationship maintenance and relationship termination (Reinartz et al. 2004). Nevertheless, not much is known about the impact of social networks on consumer behavior once customers have been acquired, i.e., contingent on the initial product/brand choice decision. Extending previous research makes it likely that consumption similarities within social networks may also extend to other variables such as product usage intensity or customer-level revenue. Yet, no study has until now investigated whether such thinking is indeed appropriate.

Our analysis intends to provide a contribution in this area. Using the customer base of a mobile phone provider, we apply approaches from spatial statistics to analyze the distribution of customer-level revenue within a social network. For this, we rely on two samples consisting of 6,681 and 19,668 actors linked by 19,885 and 25,799 ties that combine information about social relationships between customers (obtained through the analysis of call history information) with revenue data. Our analysis provides an indication for the existence of significant and substantial positive network autocorrelation in customer-level revenue. This result remains present when accounting for a potential impact of actor homophily (i.e., shared basic or socio-demographics) and when analyzing similarities in service plan choice instead of actual customer-level revenue. Positive network autocorrelation implies that high (low) revenue customers tend to be primarily related to (i.e., have strong social ties with) other high (low) revenue clients. Additionally, we find that similar but significantly smaller effects can be observed when looking at geographical (vs. communicative) proximity among customers. Hence, an approximation of social (communicative) relationships among customers with physical proximity leads to a significant underestimation of the extent of similarity and the degree of influence within the same social network.

2 Hypothesis development

Social network analysis is part of a research stream, which assumes that people cannot be reduced to a set of individual-level attributes (e.g., attitudes and socio-demographic information) and analyzed in isolation but instead need to be considered as part of a social environment whose structures, constraints, and opportunities influence their behavior (e.g., Mizruchi 1994). Its foundations can be traced back to areas such as sociometry, which was developed by Moreno shortly after World War I (see Moreno 1941 for an introduction), and structural sociology. Over the past few decades, social network analysis has been applied regularly in the marketing discipline. Exemplary applications include analyses of customer referral behavior (e.g., Money et al. 1998), investigations of the impact of word-of-mouth communication (Goldenberg et al. 2007), information sharing in new product development processes (e.g., Ahuja et al. 2003; van den Bulte and Moenaert 1998), knowledge diffusion (e.g., Levin and Cross 2004; Singh 2005), and journal citation analysis (e.g., Baumgartner and Pieters 2003). More recently, the rise of online communities such as Facebook or My Space has further increased the interest in social network analysis, specifically in network dynamics and social capital production (e.g., Mathwick et al. 2008).

Consistently, these articles have found strong support for the impact of social networks on consumer behavior and consumption patterns. In one of the first papers, for example, Reingen et al. (1984) investigated brand choice among circles of friends and found support for significant brand congruency among strong tie relationships. In our analysis, we assume that if people within the same social network tend to have similar preferences and consumption patterns for certain products or services (e.g., for a certain brand within a product category), it can be expected that they will also have similar revenue potential for the company who offers these products/services (e.g., the manufacturer of this brand). This leads to the following hypothesis:

> *H_1: Customer-level revenue within a social network shows a significant degree of positive network autocorrelation.*[1] *High (low) revenue customers are primarily directly connected to other high (low) revenue customers.*

A key challenge in social network analysis is often the collection of information about social relationships between individuals. Several studies have, for example, shown that information about the existence and strength of ties obtained by questioning respondents can be subject to significant measurement error. Marin (2004) highlights that respondents tend to forget to mention a significant number of their social relationships when being asked to provide a list of people with whom they discuss important matters. Additionally, the people named are not a representative subset of all relevant relationships since a person is more likely to be elicited when (a) s/he shares a stronger tie with the respondent and/or (b) s/he is connected to a greater number of people within the social network. In addition to this unconscious misrepresentation of truth, respondents may also purposely provide incorrect data. As shown by Feld and Carter (2002), respondents tend to over-report interactions with attractive people and under-report interactions with those who are unattractive.

Due to such bias and the substantial effort involved in collecting network data as well as the limited availability of such information in certain situations, many researchers have used proxy measures for the existence and strength of social (communicative) relations, the most common being geographical (spatial) proximity (e.g., Garber et al. 2004; Manchanda et al. 2008). Conditional on the existence of true autocorrelation (Manski 1993), we assume that the use of geographical

[1] Autocorrelation refers to the correlation of a variable (e.g., customer-level revenue) with itself. It is a term frequently used in the context of time-series analysis where it describes the correlation between two values of the same variable measured at different points in time (i.e., temporal autocorrelation). It is, however, also used in spatial statistics to represent the correlation between two values of the same variable measured at different locations (i.e., spatial autocorrelation) and social network analysis.

proximity to measure social network effects is likely to lead to an underestimation of the true extent of such effects. While people who are physically close to each other may have a higher chance of being part of the same social network, spatial proximity is likely to both overstate the intensity of relationships among people who are physically close and omit a substantial share of social contacts that may be equally emotionally close although further apart. We, therefore, state the following hypothesis:

> *H_2: Approximating communicative proximity by spatial proximity leads to a significant underestimation of the extent of positive network autocorrelation in customer-level revenue.*

3 Data collection

A main limitation of many traditional social network studies is their reliance on samples of very limited size, which are often a consequence of the highly cumbersome and resource-consuming data collection procedure usually recommended for such studies (Reingen and Kernan 1986). To overcome this weakness, several recent studies have relied on the use of large-scale databases in the context of social network analysis. Exemplary applications in the marketing area include the work of Hill et al. (2006) and Goldenberg et al. (2009) in the area of new product adoption or Trusov et al. (2009) on the comparison of the effectiveness of word-of-mouth and traditional advertising. Our analysis integrates this stream of research as it uses the call history database of a mobile phone provider to reconstruct the social network of a set of customers.

This database-driven approach to social network analysis has at least three advantages compared to more traditional techniques: First, it enables us to work with sample sizes that are one to two orders of magnitude larger than those used in many traditional studies. Second, it relies on actual communication patterns to identify social relationships and does therefore not suffer from problems related to stated information about the existence and strength of ties. Finally, it allows us to combine social network information with revenue data, which is a necessary condition to test our hypotheses.

Specifically, we collaborated with a mobile phone provider in Europe who granted us access to its customer and call history database, out of which we collected two samples—see Appendix for additional details. This mobile phone provider sells a wide range of different service plans, which represent contracts of a fixed duration that give the user access to a monthly allowance of calling minutes and text messages in exchange for a fixed monthly service fee (e.g., x minutes of calling time and y text messages for \$$z$ per month). Any communication in excess of the allowance is charged at a marginal price. In the logic of Lambrecht et al. (2007), these service plans represent three-part tariffs, which imply that there is no perfect correlation between the revenue generated by a customer and the total duration of outbound calls.

Our final sample consists of 6,681 (19,668) actors who are linked by 19,885 (25,799) call relationships or ties for sample A (sample B), respectively. The total number of friends per actor ranges from a minimum of 1 to a maximum of 42 for sample A and 82 for sample B with a mode of 1 in both cases. On average, every actor in sample A (sample B) is linked to 2.98 (1.31) other actors.

There are two points with respect to our data collection method that need to be highlighted: First, our approach only identifies friends who are customers of the same mobile phone provider. Hence, all people who have a contract with a different mobile phone provider do not appear as friends in our analysis and nor do friends who are not contacted regularly by mobile phone calls. Second, our sampling approach is likely to be subject to an endogenous network formation bias as the behavior under investigation (i.e., mobile phone calls) is used to determine inclusion in the network (see Nair et al. 2010 for a discussion and potential remedy). Although we tried to test the robustness of our findings with respect to this bias by replicating our analysis using service plan choice instead of customer-level revenue (see Section 5), only the use of truly exogenous variables (e.g., other types of communication besides mobile phone calls) to determine inclusion into the network could have truly resolved this issue.

For each actor in our two samples, we then calculated average monthly revenue by dividing the total revenue generated by the customer over his/her lifetime by the number of months in which the customer had maintained a relationship with the company. In order to account for the confidentiality of revenue data, all monthly customer-level revenue information in our sample has been normalized (i.e., divided by the overall average). The resulting normalized customer-level revenue has a median of 0.81 (0.82) and a variance of 0.56 (0.80) for sample A (sample B), respectively (due to normalization the mean is equal to 1.00 in both cases). As indicated by the fact that mean revenue is larger than median revenue, the distribution of revenue is positively (right) skewed. To account for this lack of normal distribution in customer-level revenue data, all subsequent analyses have been performed based on a logarithmic transformation instead of actual revenue information.

4 Model estimation and results

For our analysis, we calculated two different proximity measures for all pairs of actors *i* and *j* in our samples. The first one (communicative proximity) indicates the percentage of total call duration of actor *i* that is devoted to actor *j*. The second one (inverse geographical proximity) is a measure of spatial distance between actors *i* and *j* calculated based on latitude and longitude information on ZIP-code level.[2] Subsequently, we determined the degree of network autocorrelation for both distance measures by applying two well-known measures from spatial statistics: Moran's *I* (Moran 1950) and Geary's *C* (Geary 1954). Moran's *I* and Geary's *C* are two measures of spatial autocorrelation for metric variables that have been used occasionally in the marketing discipline—for example by Bronnenberg and Mahajan (2001) in their analysis of spatial dependence in retailer behavior. They have,

[2] We calculated the distance between all pairs of actors and subsequently normalized those distances by dividing them by the maximum sample distance. Inverse geographical proximity was then defined as 1− normalized distance.

however, to the best of our knowledge, never been applied in the context of social network analysis. Moran's I is defined as follows:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} \left(X_i - \overline{X}\right) \left(X_j - \overline{X}\right)}{\sum_i \left(X_i - \overline{X}\right)^2} \tag{1}$$

where N is the number of units (actors) indexed by i and j, X is the focal variable, $\overline{X}$ is the mean of X, and w_{ij} is a measure of proximity between actors i and j. Moran's I is essentially a standardized form of the weighted cross product $w_{ij}\left(X_i - \overline{X}\right) \times \left(X_j - \overline{X}\right)$ over all pairs of connected actors and comparable to a Pearson correlation coefficient in the sense that it is bounded between −1 and +1 with larger values, indicating a higher degree of network autocorrelation. The expectation of Moran's I under the absence of autocorrelation is $-1/(N-1)$ (approximately equal to zero for large samples) so that values larger (smaller) than this threshold indicate positive (negative) network autocorrelation.

Geary's C, which is an alternative measure of network autocorrelation, is inversely related to Moran's I and defined as follows:

$$C = \frac{1}{2} \frac{N-1}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} \left(X_i - X_j\right)^2}{\sum_i \left(X_i - \overline{X}\right)^2} \tag{2}$$

Similar to Moran's I, Geary's C can be considered as a standardized weighted cross product of the form $w_{ij}\left(X_i - X_j\right)^2$ over all pairs of connected actors. Geary's C can reach values between 0 and 2, where 1 indicates no autocorrelation and values between 0 and 1 positive network autocorrelation. In comparison to Moran's I, which measures global autocorrelation (relative to the overall mean $\overline{X}$), Geary's C is more sensitive to local autocorrelation (between two connected actors X_i and X_j). All calculations in our analysis have been performed using the spdep package (Version 0.4-2) in the R computation environment (Build 2.6.2).[3]

Table 1 summarizes the estimated values of Moran's I and Geary's C (actual value of the test statistic and its variance and expectation under randomization, i.e., random permutations of the focal variable X for the given weighting scheme) for the two different samples and proximity measures. It can be seen that all values of Moran's I are significantly positive, and all values of Geary's C are significantly smaller than 1. In line with H_1, this indicates the presence of a significant degree of positive network autocorrelation. In addition, Moran's I based on inverse geographical proximity is always smaller than that for communicative proximity and Geary's C based on inverse geographical proximity is either larger than (sample B) or equal to (sample A) that for communicative proximity. Hence, while the approximation of communicative proximity by spatial proximity leads to results that

[3] Specifically, we used the following functions: *moran.test* and *geary.test* to calculate Moran's I and Geary's C; *lm.morantest* to calculate Moran's I for regression residuals and *errorsarlm* to estimate the spatial simultaneous autoregressive error model in the context of the first robustness check; and *joincount.test* to calculate join-count statistics in the context of the second robustness check.

are directionally similar, the extent of positive network autocorrelation is consistently underestimated (H_2).

The difference in the degree of network autocorrelation, which we observe between the two proximity measures, is likely to be influenced by two different effects: Approximating social relationships by geographical proximity may falsely assume that (a) people who are physically close are also close friends and (b) people who are living far apart from each other only maintain weak social relationships. To understand the relative importance of these two different effects better, we calculated a third proximity measure, referred to as "mixed" in Table 1. This measure is equal to communicative proximity for all cases where two actors share the same ZIP code (resulting in a geographical proximity of 0 and an inverse geographical proximity of 1) and equal to inverse geographical proximity in all other cases. Mixed proximity therefore corrects inverse geographical proximity for the fact that people who are physically close may not necessarily be close friends. As can be seen in Table 1, mixed proximity leads to values for Moran's I and Geary's C that are between those obtained for communicative and inverse geographical proximity in three out of four cases. It appears that 20–40% in the difference between communicative and inverse geographical proximity can be explained by the overestimation of social influence for individuals who are physically close.

5 Robustness checks

In order to test the stability of our findings, we performed two additional analyses as a robustness check: First, we determined Moran's I for both samples while controlling for a potential impact of actor homophily, and second, we analyzed social network effects with respect to service plan choice instead of actual customer-level revenue within sample B.

Regarding the first robustness check, controlling for the impact of actor homophily, it has been discussed in the literature that social influence effects may exist because people generally prefer to interact with others who are similar to themselves (e.g., Reagans 2005), so that similarities in consumption patterns among people with similar profiles tend to extend to the social network. House and Mortimer (1990) highlight that social stratification (e.g., by age, by occupational position) is likely to have a significant impact on social network composition, and marketing literature has regularly shown that the same variables have a significant impact on consumer behavior. Also, a review article of McPherson et al. (2001) provides substantial evidence for the presence of homophily regarding variables such as race/ethnicity, sex/gender, age, religion, and education/occupation/social class.

To test the stability of our results with respect to actor homophily, we obtained information on basic demographics (i.e., age, gender, marital status) and socio-demographics for each actor in our samples.[4] For the latter, we relied on the Mosaic

[4] While this was possible for virtually all actors within sample B (only 55 out of 19,668 actors showed missing values with respect to the age variable), we were only able to obtain such information for roughly 42% of the 6,681 actors within sample A. The results for the latter case therefore need to be interpreted with caution.

Table 1 Degree of spatial and network autocorrelation in customer-level revenue

Sample	Proximity measure	Geary's C			Moran's I			Simultaneous autoregressive error model (λ)
		Value	Variance	Expectation under randomization	Value	Variance	Expectation under randomization	
Baseline model								
Sample A	Communicative	0.6382	0.000827	1.0000	0.5241	0.000078	-0.0001	
	Mixed	0.6214	0.001354	1.0000	0.5042	0.000090	-0.0001	
	Inverse geographical	0.6000	0.002943	1.0000	0.4149	0.000176	-0.0001	
Sample B	Communicative	0.8215	0.002901	1.0000	0.4118	0.000122	-0.0001	
	Mixed	0.8296	0.003241	1.0000	0.3293	0.000091	-0.0001	
	Inverse geographical	0.7966	0.006108	1.0000	0.1717	0.000140	-0.0001	
Robustness check – Control for impact of actor homophily								
Sample A	Communicative				0.4823	0.000208	-0.0032	0.8406
	Mixed				0.4624	0.000233	-0.0025	0.2864
	Inverse geographical				0.3780	0.000431	-0.0021	0.2107
Sample B	Communicative				0.3374	0.000359	-0.0027	0.4044
	Mixed				0.2447	0.000223	-0.0017	0.2938
	Inverse geographical				0.1269	0.000305	-0.0002	0.1872

segmentation system. Mosaic is a geo-demographic segmentation system developed by the market research firm Experian that groups customers into different types (referred to as Mosaic codes) based on a series of segmentation variables such as property type, household composition, length of residency, employment, and income (see Farr and Webber 2001 for more details on the segmentation logic applied). Mosaic codes have previously been used in marketing literature as a classification tool, for example to profile advertising complainants (Crosier and Erdogan 2001; Volkov et al. 2005) or to identify people particularly likely to suffer from certain diseases in the context of social marketing campaigns (Powell et al. 2007).[5]

To control for a potential impact of actor homophily in the context of our study we performed two different analyses: First, we estimated a regression model using the log of average customer-level revenue as the dependent variable and age, gender, marital status, and Mosaic code as independent variables. We then determined Moran's I in both samples based on the regression residuals instead of actual customer-level revenue. It turns out that basic and socio-demographics are only weak predictors of customer-level revenue (adjusted R^2 of approximately 3.2% for sample A and 7.8% for sample B). This is consistent with previous research, which has shown that such variables are only weakly correlated with monetary measures of customer attractiveness, such as future customer profitability or customer lifetime value (Campbell and Frei 2004). As a consequence, taking account of the potential impact of demographic information only leads to a small reduction in Moran's I (about 10% for sample A and 20% for sample B; see Table 1).

It may, however, be possible that the two-step approach we applied to determine these corrected values for Moran's I (i.e., first estimating a regression model and then using the regression residuals to determine Moran's I) did result in inefficient or biased estimates. To rule out this explanation, we performed a second analysis using a simultaneous autoregressive error model (Ord 1975). Specifically we estimated a model of the form:

$$y = X\beta + u \,\text{with}\, u = \lambda W u + \varepsilon \tag{3}$$

where y is the focal independent variable (i.e., log of average customer-level revenue), X is a matrix of actor attributes (i.e., age, gender, marital status, and Mosaic code), W is the proximity matrix (i.e., the matrix of all measures of proximity between actors i and j w_{ij}), and u and ε are error terms. A simultaneous autoregressive error model takes account of the potential impact of actor attributes on the independent variable (reflected in the parameter vector β) while at the same time reflecting network autocorrelation in the regression residuals. The parameter λ hereby represents the degree of network autocorrelation that remains present when accounting for actor attributes, and the model collapses to the standard regression model in case of λ=0. As shown in Table 1, the model estimation results in substantial values of λ for communicative proximity (around 0.80 for sample A and 0.40 for sample B) and in substantial but significantly smaller values for inverse geographical proximity (around 0.20 for both samples). Combined, this shows that our findings remain robust, even when accounting for a potential impact of actor homophily.

[5] See http://www.appliedgeographic.com/mosaic.html for additional details on the Mosaic typology.

With respect to the second robustness check, social network effects in service plan choice, it is conceivable that our analysis suffers from a bias as both information on network ties and actor attributes (customer-level revenue) stem from the same event (mobile phone calls). As highlighted above, the three-part tariff structure of the mobile phone operator leads to the fact that there is no perfect correlation between the revenue generated by a customer and the total duration of outbound calls. Nevertheless, given that outbound calls that exceed the monthly allowance are charged at a marginal (per minute) price, some correlation between both variables is to be expected.

To test for a potential bias introduced by this correlation, we replicated our analysis using service plan choice instead of customer-level revenue as dependent variable. Each service plan represents a certain contract of a fixed duration that gives the user access to a monthly allowance of calling minutes and text messages in exchange for a fixed monthly service fee. All customers who subscribe to the same service plan pay the same fixed monthly service fee, independent from their actual mobile phone usage. We therefore expect only a weak correlation between actual mobile phone usage and the service plan choice for any given user, except for the fact that the initial service plan choice is likely to be a function of a customer's anticipated mobile phone consumption.

We obtained service plan information for each of the 19,668 actors within sample B, leading to a total of 331 different service plans. Out of these 331 different options, 15 represent more than 50% of all choices. We subsequently computed join-count statistics (Moran 1948) for both proximity measures and 15 service plans. Join-count statistics are the simplest measure of spatial autocorrelation and are used for binary variables. If the two values of the binary variable are referred to as "black" and "white" there are three different types of unordered "joins" (i.e., joining areas or neighborhood relationships) possible, namely black–black, black–white, and white–white. Join-counts are counts of the number of different joins in the network under investigation. These join-counts can subsequently be compared to their expectation under the null hypothesis of no autocorrelation to test for the presence of significant autocorrelation.

Our analysis indicates substantial network autocorrelation in service plan choice for both communicative and inverse geographical proximity.[6] Yet, while for communicative proximity the join-count statistics are significant for all 15 service plans, we only observe significant network autocorrelation in eight out of 15 cases for inverse geographical proximity. Combined, this shows that service plan choice shows a significant degree of positive network autocorrelation and that approximating communicative proximity by spatial proximity leads to a significant underestimation of the extent of positive network autocorrelation in service plan choice—consistent with our previous results.[7]

[6] Details on this analysis can be obtained from the author on request.

[7] The three-part tariff structure of the mobile phone operator induces a relationship between customer-level revenue and service plan choice. Mean customer-level revenue for all clients within the same service plan ranges from 0.23 to 1.72 for the 15 service plans analyzed. Nevertheless, due to substantial mobile phone usage outside of the monthly service plan allowance, the standard deviations of the mean are significant (between 0.23 and 0.98). This leads to the fact that although some service plans are associated with strictly larger or smaller customer-level revenue than others, many service plans overlap in terms of customer-level revenue. Service plan choice and customer-level revenue are therefore two distinct measures of post-acquisition customer behavior.

6 Limitations and areas of future research

Our analysis is likely to suffer from the two limitations that future studies should address. This would strengthen our key results that customer-level revenue within a social network shows a significant degree of positive network autocorrelation and that approximating communicative proximity by spatial proximity leads to a significant underestimation of these effects.

First, one could argue that our study does not truly address the impact of geography on social networks as our measure of geographical distance based on latitude and longitude information on ZIP-code level is not sufficiently precise, and more generally, the key purpose of telecommunication services is to be able to get in touch with people who are not geographically close. Although our sample includes more than 2,800 different ZIP codes, each one still summarizes either several smaller villages or a larger area in one big city. Unfortunately, the mobile phone provider was not able to provide us with geographical information on street level as the combination of detailed location data and call history information would have resulted in legal problems. Additionally, although about 60% of all calling relationships in sample A and 30% in sample B live in the same ZIP-code area (resulting in a physical distance of zero), the average distance between two communication partners is still considerable—approximately 50 km in both samples. Nevertheless, we would like to highlight that there is no significant correlation between communicative and geographical proximity in both samples (Pearson correlation coefficient of −0.02 in sample A and 0.07 in sample B). It does therefore not appear that people who are called more often tend to be further apart.

Second, as with virtually all social network studies, our findings may suffer from contextual effects due to the possibly unique nature of the social network being studied (Manchanda et al. 2008; Manski 1993), which lead to the fact that our results may not be fully generalizable to the whole customer base of the mobile phone provider we collaborated with. Furthermore, our analysis only considers friends who are contacted by mobile phone calls and who are customers of the same mobile phone provider, which might have introduced an additional bias.

With respect to areas of future research, we consider two questions to be particularly worthwhile: First, it would be interesting to incorporate our findings into a more general model that investigates social network effects in brand choice and usage intensity (customer-level revenue) contingent on brand choice, simultaneously. Alternatively, our results could also enrich a more general stochastic model of customer-level revenue distribution, for example by extending the well-established gamma–gamma model that is usually used for such purposes (e.g., Fader et al. 2005b). Such a model could then be combined with a stochastic transaction model, such as the Pareto/NBD (Schmittlein et al. 1987) or the BG/NBD model (Fader et al. 2005a), as well as a model of the number of friends per customer, to derive an estimate of the value of a customer's social network that could subsequently be included in his/her customer lifetime value (CLV).

Second, our analysis could be extended to an investigation of social network effects in customer lifetime or churn behavior (see Krackhardt and Porter 1986; Krackhardt and Porter 1985 for a similar investigation with respect to employee turnover). While our analysis has provided an indication for positive network autocorrelation in revenue, it is unclear whether the same relationship exists with respect to loyalty/individual-level lifetime. If similar effects can be observed in this setting, it could be expected that CLV (which essentially measures revenue/profit over lifetime) shows an even stronger degree of network autocorrelation than we observe with respect to revenue. Another question of interest in that context would be to what extent social effects play a role in individual-level churn decisions. While it has been discussed that social contagion plays a role in new product adoption, it would be interesting to investigate whether the same holds true for disadoption and/or churn, that is, whether customers are more likely to leave a company if one or several of their friends has previously left.

Appendix: Sampling process

Sample A

We started the creation of sample A by taking a random sample from the customer database of the mobile phone provider. Since every customer can be uniquely identified by his/her mobile phone number, we generated a list of 150,000 random numbers and matched it to the customer database.[8] This resulted in a random sample of 363 customers. For each of these 363 customers, we downloaded information about all outgoing calls made (phone number called and duration of call) over a 3-month time period (March 1 to May 31). We then calculated the total number of minutes any number had been called and expressed it as a percentage of total call duration. Any number which represented at least 1% of total call duration was subsequently considered as a potential friend of the calling customer.[9] We then matched this list of mobile phone numbers back to the customer database, resulting in 747 customers of the mobile phone provider that could be considered as friends of at least one of the initial 363 customers. Following the same procedure again, we subsequently identified another 2,639 customers (either friends of the initial 363 and/or the 747 customers) and 6,966 customers (either friends of the initial 363 and/or 747 and/or 2,639 customers). In the resulting list of 10,715 customers (363+747+2,639+6,966), 2,710 customers were deleted due to double counting and 292 because they had been acquired after March 1 and, hence, only had incomplete call history information. Out of the remaining 7,713 customers, 7,055 could be matched

[8] The mobile phone company we collaborated with is based in Europe, where the predominant billing scheme is Calling Party Pays. This implies that the mobile subscriber does not pay for incoming calls but instead the calling party pays for those calls. In order to notify the calling customer that s/he has called a number for which there will be a different tariff, mobile numbers in Europe are usually dedicated to specific blocks. This made it straightforward to generate a set of random numbers which had a reasonably high chance of corresponding with actual mobile phone numbers.

[9] It was necessary to define a cut-off in terms of call duration to eliminate numbers from our analysis that have been called only rarely and that, hence, are unlikely to represent true friends. In line with the well-established 80/20 law, we assumed that friends should account for 80% of total call duration. We subsequently tested a range of potential thresholds between 1% and 5% to identify to what extent they fulfilled this criterion. While a 1% cut-off resulted in friends accounting for 84% of total call duration, a 2% cut-off would have resulted in 73% and a 3% cut-off in 64%. Based on these results, we decided to apply a 1% threshold.

to a second database containing revenue information and, out of those, 6,681 to a third one containing postcode data. This resulted in a final network consisting of 6,681 actors, who were linked by 19,885 call relationships or ties, with a density of 0.0005628. For our analysis, we transformed all directed relationships into undirected ones (i.e., arcs into edges) to obtain a symmetrical adjacency matrix as the underlying event (mobile phone calls) is by nature a reciprocal relationship.

Sample B

Similar to the approach taken for sample A, the creation of sample B started by matching a set of random numbers (1.25 million) to the customer database of the mobile phone provider, resulting in a random sample of 4,163 customers. Deleting 424 customers that had been acquired after August 1 led to 3,739 customers, for which we downloaded information about all incoming and outgoing calls made (phone number and duration of call) over a 2-month time period (August 1 to September 30). This resulted in 12,939 customers of the mobile phone provider that called at least one of the initial 3,739 customers and 12,853 that were called by at least one of them. In the resulting list of 29,531 customers (3,739+12,939+12,853), 9,689 customers were deleted due to double counting. Out of the remaining 19,842 customers, 19,826 could be matched to a second database containing revenue information and, out of those, 19,668 to a third one containing postcode data. This resulted in a final network consisting of 19,668 actors, who were linked by 25,799 call relationships or ties, with a density of 0.0000851. As above, we transformed all directed relationships into undirected ones (i.e., arcs into edges) to obtain a symmetrical adjacency matrix as the underlying event (mobile phone calls) is by nature a reciprocal relationship.

References

Ahuja, M. K., Galletta, D. F., & Carley, K. M. (2003). Individual centrality and performance in virtual R&D groups: An empirical study. *Management Science, 49*(1), 21–38.

Baumgartner, H., & Pieters, R. (2003). The structural influence of marketing journals: A citation analysis of the discipline and its subareas over time. *Journal of Marketing, 67*(2), 123–139.

Bronnenberg, B. J., & Mahajan, V. (2001). Unobserved retailer behavior in multimarket data: Joint spatial dependence in market shares and promotion variables. *Marketing Science, 20*(3), 284–299.

Campbell, D., & Frei, F. (2004). The persistence of customer profitability: Empirical evidence and implications from a financial services firm. *Journal of Service Research, 7*(2), 107–123.

Crosier, K., & Erdogan, B. Z. (2001). Advertising complainants: Who and where are they? *Journal of Marketing Communications, 7*(2), 109–120.

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005a). Counting your customers the easy way: An alternative to the Pareto/NBD model. *Marketing Science, 24*(2), 275–284.

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005b). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research, 42*(4), 415–430.

Farr, M., & Webber, R. (2001). MOSAIC: From an area classification system to individual classification. *Journal of Targeting, Measurement and Analysis for Marketing, 10*(1), 55–65.

Feld, S. L., & Carter, W. C. (2002). Detecting measurement bias in respondent reports of personal networks. *Social Networks, 24*(4), 365–383.

Frenzen, J. K., & Davis, H. L. (1990). Purchasing behavior in embedded markets. *Journal of Consumer Research, 17*(1), 1–12.

Garber, T., Goldenberg, J., Libai, B., & Muller, E. (2004). From density to destiny: Using spatial dimension of sales data for early prediction of new product success. *Marketing Science, 23*(3), 419–428.

Geary, R. C. (1954). The continguity ratio and statistical mapping. *The Incorporated Statistician, 5*(3), 115–145.

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), 211–223.

Goldenberg, J., Libai, B., Moldovan, S., & Muller, E. (2007). The NPV of bad news. *International Journal of Research in Marketing, 24*(3), 186–200.

Goldenberg, J., Han, S., Lehmann, D. R., & Hong, J. W. (2009). The role of hubs in the adoption process. *Journal of Marketing, 73*(2), 1–13.

Hill, S., Provost, F., & Volinsky, C. (2006). Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science, 21*(2), 256–276.

House, J. S., & Mortimer, J. (1990). Social structure and the individual: Emerging themes and new directions. *Social Psychology Quarterly, 53*(2), 71–80.

Kelman, H. C. (1961). Processes of opinion change. *Public opinion Quarterly, 25*(1), 57–78.

Krackhardt, D., & Porter, L. W. (1985). When friends leave: A structural analysis of the relationship between turnover and stayers' attitudes. *Administrative science quarterly, 30*(2), 242–61.

Krackhardt, D., & Porter, L. W. (1986). The snowball effect: Turnover embedded in communication networks. *Journal of Applied Psychology, 71*(1), 50–55.

Lambrecht, A., Seim, K., & Skiera, B. (2007). Does uncertainty matter? Consumer behavior under three-part tariffs. *Marketing Science, 26*(5), 698–710.

Levin, D. Z., & Cross, R. (2004). The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer. *Management Science, 50*(11), 1477–1490.

Manchanda, P., Xie, Y., & Youn, N. (2008). The role of targeted communication and contagion in product adoption. *Marketing Science, 27*(6), 961–976.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies, 60*(3), 531–542.

Marin, A. (2004). Are respondents more likely to list alters with certain characteristics? Implications for name generator data. *Social Networks, 26*(4), 289–307.

Mathwick, C., Wiertz, C., & de Ruyter, K. (2008). Social capital production in a virtual P3 community. *Journal of Consumer Research, 34*(6), 832–849.

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

Mizruchi, M. S. (1994). Social network analysis: Recent achievements and current controversies. *Acta Sociologica, 37*(4), 329–343.

Money, R. B., Gilly, M. C., & Graham, J. L. (1998). Explorations of national culture and word-of-mouth referral behavior in the purchase of industrial services in the United States and Japan. *Journal of Marketing, 62*(4), 76–87.

Moran, P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society: Series B (Methodological), 10*(2), 243–251.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37*(1/2), 17–23.

Moreno, J. L. (1941). Foundations of sociometry: An introduction. *Sociometry, 4*(1), 15–35.

Nair, H. S., Manchanda, P, & Bhatia, T. (2010). Asymmetric social interactions in physician prescription behavior: the role of opinion leaders. Journal of Marketing Research (in press).

Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association, 70*(349), 120–126.

Powell, J., Tapp, A., & Sparks, E. (2007). Social marketing in action—Geodemographics, alcoholic liver disease and heavy episodic drinking in Great Britain. *International Journal of Nonprofit and Voluntary Sector Marketing, 12*(3), 177–187.

Reagans, R. (2005). Preferences, identity and competition: Predicting tie strength from demographic data. *Management Science, 51*(9), 1374–1383.

Reinartz, W., Krafft, M., & Hoyer, W. D. (2004). The customer relationship management process: Its measurement and impact on performance. *Journal of Marketing Research, 41*(3), 293–305.

Reingen, P. H., & Kernan, J. B. (1986). Analysis of referral networks in marketing: Methods and illustration. *Journal of Marketing Research, 23*(4), 370–378.

Reingen, P. H., Foster, B. L., Brown, J. J., & Seidman, S. B. (1984). Brand congruence in interpersonal relations: A social network analysis. *Journal of Consumer Research, 11*(3), 771–783.

Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting your customers: Who are they and what will they do next? *Management Science, 33*(1), 1–24.

Singh, J. (2005). Collaborative networks as determinants of knowledge diffusion patterns. *Management Science, 51*(5), 756–770.

Trusov, M., Bucklin, R. E., & Pauwels, K. (2009). Effects of word-of-mouth versus traditional marketing: Findings from an Internet social networking site. *Journal of Marketing, 73*(5), 90–102.

van den Bulte, C., & Moenaert, R. K. (1998). The effects of R&D team co-location on communication patterns among R&D, marketing and manufacturing. *Management Science, 44*(11, Part 2 of 2), S1–S18.

Volkov, M., Harker, D., & Harker, M. (2005). Who's complaining? Using MOSAIC to identify the profile of complainants. *Marketing Intelligence & Planning, 23*(3), 296–312.