Peer influences in air travel behavior: A social autoregressive analysis

Gregory S. Macfarlane, Thomas A. Wall, Bingling Zhang, Kari E. Watkins

Abstract:

Social networks can influence travel in two distinct ways: first, people often make trips to see people they know through work or life (exogenous effects); or they make trips that are similar to those taken by friends or family (endogenous effects). In spite of this, traditional models have not always made use of social network data, because information on social networks is both difficult to collect and challenging to implement in models. In this paper, we appropriate techniques developed for *spatial* data (specifically, a spatial Durbin count model) to construct a trip generation model for *socially* linked observations. We train this model on data collected through a web survey that interacts with the Facebook application programming interface (API) to capture respondents’ “friends” lists. We construct a set of “social weights” matrices that captures the social tie-strength between observations. A matrix that weights two observations by the number of their mutual friends provides the maximum model likelihood; this model further suggests that a non-social model may result in erroneous interpretations of the link between demographic characteristics and trip generation.

*Keywords*: autoregression, social networks, travel behavior, spatial Durbin model, Facebook

# Introduction

The goal of a modern activity-based travel model is to accurately represent travel behavior choices on a microscopic level (Bhat & Koppelman 2003). As researchers have pursued this goal, they have sought to incorporate individual behavioral patterns of various types. For example, researchers have built models that allow different classes of people to place different values on goods (Walker & Ben-Akiva 2002), that capture idiosyncratic preferences for a particular lifestyle (Cao et al. 2006), or that allow multiple decisions to be made either sequentially or simultaneously (Bhat & Sen 2006; Pinjari et al. 2011).

Another dimension of individual behavior that is receiving increasing attention from transportation researchers is the role that peer or social networks play in travel choices. These influences may come in several ways:

Direct: each individual chooses activities or destinations where their social connections are, or where they will be.

Indirect: an individual’s social connections influence their choice of destinations or activities.

Endogenous: individuals construct their social networks with others of shared interest or experience.

The distinction between these elements can be subtle. Imagine two individualswho are socially connected to each other. If these individuals meet for dinner, this is a direct influence, because each person’s choice of activity and destination is dependent on the other person being there. On the other hand if one recommends a restaurant to the other, this is an indirect influence because the restaurant entered one’s choice set through the other’s experience and not his own. Finally, if the two know each other through another shared destination or interest (e.g. workplace, neighborhood, church), the restaurant in question may have entered their choice set by being near one of these destinations.

These elements are important to capture in activity-based models for several reasons. First, they are a natural element of every day behavior. But more importantly, an activity model that does not consider them might produce spurious or unreliable predictions, because standard econometric assumptions that permit consistency and validity may not apply. Despite this importance, models that incorporate social influences are rare for two reasons: because social network data has been difficult to collect, and because there has not been a convenient econometric framework to apply.

However, both of these situations are changing. Usage of online social networking sites and other social media continues to grow, and contains social connection and preference data for millions of individuals. On the econometric front, the spatial econometrics literature in the last twenty-plus years (Anselin 2003; Anselin & Arribas-bel 2011) enables the view of spatial interaction as a special case of social interaction, or as a proxy that may be more easily observed in data.

In this paper, we develop and apply a technique that uses spatial data to analyze airline travel (trip frequency) of individuals in a common social network. To do this, we construct a series of social weights matrices representing potential relationship strength between members of this network, and incorporate these matrices in a social Durbin model capturing direct, indirect, and endogenous effects.

The paper is organized as follows: Section 2 presents the literature establishing spatial/social econometrics as well as previous studies establishing the effect of social networks on travel behavior. Section 3 presents our model and the training data that we collected from a sample of undergraduate students (specifically those in the fraternity and sorority, or "Greek," system), with results from the model presented in Section 4. Section 5 concludes the paper with implications for future practice and research.

# Social Networks and Spatial (Econometrics? Influence?)

## Social Networks, Online Observation, and Facebook

The concept of a social network provides a construct in which the structural and functional dimensions of individuals’ social relationships ([O'Reilly 1988](#_ENREF_4)) can be systematically examined for social influence, or other characteristics of interest. Social network structure, which refers to the individuals and linkages comprising one's social network ([Due, Holstein et al. 1999](#_ENREF_1)) can be analyzed either through egocentric network analysis (focused on a single individual and his relationships with others) or whole network analysis (focused comprehensively on the relationships among all members in a network). Social network function clarifies the nature of, and interpersonal interaction among, the linkages within the social network structure; for example, "social support, relational strain, or social anchorage," among others. ([Due, Holstein et al. 1999](#_ENREF_1)) In the context of this study, one important dimension of social network function could be the level of influence associated with an individual connection between two individuals, or among many connections within a group of individuals.

Data collection, however, complicates this type of study. The predominant method employed by previous studies (for example, *(27–29)*) to construct models of individuals’ social networks, called “name generators,” uses sets of interview questions designed to solicit information from an individual regarding whom he considers as part of his social network (structure) and the role or importance of that individual (function). The name generator approach can become problematic, however, when respondents have difficulty in recalling information regarding their social networks *(15)*, thus providing information to the researchers that is unknowingly incomplete. In addition, the name-generator approach can be time consuming, with data collection interviews frequently taking several hours for a single interview *(30)* to construct a model of one individual's egocentric social network.

Online social networking websites, such as Facebook and Twitter, provide an alternative source for social network data, the collection of which allows for the delineation of naturally developing social networks with unambiguous boundaries (thus overcoming respondent recollection issues) *(14, 15)*. Additionally, as such data is stored in a centralized computer database and in a format that is consistent across all users, social network data can be collected in seconds and with greater structural consistency. Thus, data collection is far less time consuming and data quality and consistency is potentially improved as compared to the name generator approach *(14, 15)*.

In this study, online social network data collection is focused exclusively on users of the social networking site, Facebook. With 864 million daily active users, and 1.12 billion monthly active users globally as of September 2014, Facebook is among the world's largest social networking sites ([Facebook 2015](#_ENREF_2)) (by comparison, Twitter has 284 million monthly active users across the globe as of October 2014). ([Twitter 2015](#_ENREF_6)). Facebook enables users to create online profiles (e.g., general demographic, educational, special interest, and geographical information), which is shared with other users. As users connect with other individuals, becoming 'friends,' an online model of social relationships and connections is created. ([Wall, Macfarlane et al. 2014](#_ENREF_7)). A study by the Pew Research Center found that the primary motivation for the majority of adult users of social networking sites in the United States is "staying in touch with family members," and "staying in touch with current friends and reconnecting with old friends." ([Smith 2011](#_ENREF_5)) It therefore may be reasonable to "suspect that an individual’s online social network structure represents, to some extent, an online model of that individual’s real-life social network structure." ([Wall, Macfarlane et al. 2014](#_ENREF_7))

There are also scharacteristics of the Facebook datasetthat dataset, thatLewis et al. ([2008](#_ENREF_3)) discusses five defining features of the Facebook social network model:

1. **Naturally Occurring**: Facebook data is naturally occurring when downloaded directly from Facebook (i.e. not subject to a respondent recollection ability)
2. **Whole Networks**: Facebook.com contains complete network data of a sociocentric nature, as opposed to an egocentric nature, which reflects the interconnectedness of network members
3. **Longitudinal Data:** Facebook data may reflect how relationships grow and change over time;
4. **Cultural Insight:** Facebook users can indicate tastes and values, which may be reflected in the data
5. **Relationship Strength:** Facebook can provide data on relationship strength.

The whole network aspect of Facebook data enables an analysis of interconnected or shared relationships among a population. Further, relationship strength provides some insight into the function of relationships, as may be related to the level of indirect influence across a relationship. Lastly, and very importantly, Facebook maintains locational information that can easily be translated to geospatial

## Spatial Econometrics

Spatial econometrics is built on the understanding that observations near to each other spatially are likely to exhibit correlation and interdependence, as they share common unobserved variables and influence each other. A class of models developed to solve these problems was pioneered by Anselin (1988); LeSage & Pace (2009) provide a full and modern treatment.

For a linear regression model , a necessary assumption is that the observations are independent, which is unlikely in spatial data. Tobler’s law implies that if two observations are “close” to each other, then they are likely to share common characteristics (Tobler 1970). Further, observations close to each other may actually exert an influence on each other in the way air pollution lowers the price of nearby properties (Kim & Goldsmith 2008). If correlation and interdependence are not controlled in an econometric model, then the model may return biased parameters and/or invalid hypothesis tests.

This correlation and interdependence can be accommodated with a spatial Durbin model (SDM),

where is a spatial weights matrix in which each element maps the spatial relationship between two observations and ; if and are “neighbors,” then . The parameter measures the dependence between and , and the lagged independent variable coefficients in contain the impact that attributes of have on .

Such spatial models have been applied to numerous different contexts, including transportation modeling and forecasting. In most if not all of these cases, however, spatial data is used as a proxy for an inferred social relationship. For example, Adjemian et al. (2010) observed that vehicle body types were spatially autocorrelated, with the authors inferring that neighbors may influence each other in choosing vehicle types. Similarly, Bhat et al. (2010) used spatial proximity as a means to study the social destinations of teenagers. In these and other cases, it may be more appropriate to use the social network directly, if the social network can be translated to a matrix. This is effectively the approach of Bramoullé et al. (2009) and Calvó-Armengol et al, (2009), who showed that a student's school performance and physical activity level (respectively) are socially interdependent: students with studious or athletic friends tend to be so as well.

There are two primary difficulties in creating , one econometric and one practical. From an econometric standpoint, the researcher typically must specify exogenously. Though some common practices have emerged in the spatial literature (Dubin 1998), not all of these can be readily adopted to social networks. This gap in the literature is one that we seek to fill with this study, by examining several possible matrices that we construct from social network data. The practical difficulty is actually collecting that data, though this is becoming less difficult as interaction within social networks happens through online sites such as Facebook. We discuss this practical evolution in Section 2.2.

# Methodology

We introduce social dependence into a Poisson count model that predicts the number of airline trips an individual makes as a function of his or her socioeconomic variables. The Poisson count model estimates the effect of dependent variables on the conditional mean number of trips, . We consider social interaction in this model by suggesting that (where is the row of that contains the weights for the neighbors of ). Our goal in this analysis is to recover a weighting scheme that maximizes the model likelihood, and use this model to examine the inferential error that could result from ignoring social interdependence and correlation in such model.

## Data Collection

To examine the potential for influences to travel behavior between individuals’ within an interconnected social network, a web-based survey was developed, and respondents were recruited through the Georgia Tech fraternity and sorority system (a social network that may presumably have a proportionally higher level of interconnectedness), in partnership with the Georgia Tech Dean of Students, Office of Greek Affairs. The web-based survey instrument was distributed via email to the primary points of contact for each fraternity and sorority on file at the Office of Greek Affairs. Fraternities and sororities were asked to also encouraged to solicit responses from their alumni, thus, potentially increasing the sample population of approximatley 3,300 students in Greek housing. Fraternities and sororities were incentivized with the promise of a $500 charitable contribution made on behalf of the fraternity and the sorority each with the highest participation.

The online survey instrument was adapted from Wall et al ([2014](#_ENREF_7)), and contains three principal components. For a detailed description of the variables collected, refer to Wall et al. ([2014](#_ENREF_7)). The first component collected respondents’ air travel information from the past year – this includes a list of trips taken with origin and destination airports for round-trip, multi-destination, and one-way travel. The second component collected socio-demographic information, including age, gender, income, education level, citizenship, occupation, and race. The third component asked respondents to log in to their Facebook accounts, and the survey automatically collected data contained in the respondent's profile and friends list via the Facebook application programming interface (API). This data included the hometown city and current city locations of the respondents, and also that of the respondents’ friends. The program also collected an encrypted (one way hash) version of the Facebook ID for all respondents and their friends. This ensured subject anonymity while simultaneously allowing observations regarding 1) which respondents were friends with each other, and 2) how many friends respondents have in common, 3) identifying and limiting the potential for duplicate respondents.

Because the survey program collects Facebook profile and friends list data from the perspective of the respondent's individual profile, it is limited in its data collection to "visible" user profile and friend data. That is, individuals' information that was restricted as a result of more stringent individual privacy settings would not be visible to the survey program. A consequence of this is potential for imbalance in privacy settings (and thus, data visibility) between two connections, is that friend connections may be observed as one-way in the data even though one-way friend connections are not possible on Facebook.

Facebook locational data (i.e., hometown and current city location) is not geocoded, and therefore required some post-processing prior to any analyses. Facebook location data were post-processed using R statistical programming language and ArcGIS 10. The study analyzed respondents’ friends who do not consider Atlanta, Georgia as their hometown. Friends’ hometown cities were geocoded as point locations. To associate cities as within (or outside of) the Atlanta metropolitan region, circular buffers were created (using Hartsfield-Jackson Atlanta International airport as the centroid) at 25 mile radius increments from 25 miles to 150 miles. The 6 resulting hometown buffer regions were then associated with the geocoded city locations, and thus the number of friends with hometown city locations within each buffer region were determined. Only data from respondents that completed the entire survey (i.e. the air travel diary, socio-demographic section, and the Facebook data collection) were considered for the analysis.

## Candidate Weighting Schema

The social network data we retrieved from Facebook allow us to construct four social weights matrices, which we describe in this section. To aid this presentation, we must first establish some basic notation following elementary set theory. The set of friends of an individual is defined as . The cardinality of this set is the number of friends, so therefore the quantity represents the number of friends that individuals and have in common.

### Direct Friendship

This schema defines a simple binary matrix where a cell is nonzero if and are friends. Specifically,

### Mutual Friendship

This schema defines a matrix where the cell value is the number of friends and have in common (as in Gilbert & Karahalios 2009; He et al. 2012),

### Network Overlap

This schema defines a matrix where the cell value is the relative overlap of each pair’s social network (as in Adamic & Adar 2003),

### Direct Network Overlap

This schema is similar to the Network Overlap, except that we give a bonus to pairs that are themselves friends,

By Facebook’s internal logic, each of these matrices must be symmetric: it is not possible on Facebook to have non-reciprocal friends. However, the individual user privacy settings make it possible that we would observe apparent one-way relationships through the API; indeed, this affected approximately 10% of cases. We therefore coerce this and other schema to be symmetric as appropriate.

As a final note on the schema development, behavioral inference on and estimation of the models is simplified by using a row-standardized matrix, where the rows sum to unity(LeSage & Pace 2009, p.85). We follow this convention in all cases.

# Results and Discussion

# Conclusions

References

Adamic, L. a & Adar, E., 2003. Friends and neighbors on the Web. *Social Networks*, 25(3), pp.211–230.

Adjemian, M.K., Lin, C.-Y.C. & Williams, J., 2010. Estimating spatial interdependence in automobile type choice with survey data. *Transportation Research Part A: Policy and Practice*, 44(9), pp.661–675.

Anselin, L., 1988. *Spatial Econometrics: Methods and Models (Studies in Operational Regional Science)*, Dordrecht: Kluwer.

Anselin, L., 2003. Spatial externalities, spatial multipliers, and spatial econometrics. *International Regional Science Review*, 26(2), pp.153–166.

Anselin, L. & Arribas-bel, D., 2011. *Spatial Fixed Effects and Spatial Dependence Spatial Fixed Effects and Spatial Dependence ∗*, Tempe.

Bhat, C.R. & Koppelman, F., 2003. Activity-Based Modeling of Travel Demand. In R. Hall, ed. *Handbook of Transportation Science SE - 3*. International Series in Operations Research & Management Science. Springer US, pp. 39–65.

Bhat, C.R. & Sen, S., 2006. Household vehicle type holdings and usage: an application of the multiple discrete-continuous extreme value (MDCEV) model. *Transportation Research Part B: Methodological*, 40(1), pp.35–53.

Bhat, C.R., Sener, I.N. & Eluru, N., 2010. A flexible spatially dependent discrete choice model: Formulation and application to teenagers’ weekday recreational activity participation. *Transportation Research Part B: Methodological*, 44(8-9), pp.903–921.

Bramoullé, Y., Djebbari, H. & Fortin, B., 2009. Identification of peer effects through social networks. *Journal of Econometrics*, 150(1), pp.41–55.

Calvó-Armengol, A., Patacchini, E. & Zenou, Y., 2009. Peer effects and social networks in education. *Review of Economic Studies*, 76(4), pp.1239–1267.

Cao, X., Mokhtarian, P.L. & Handy, S.L., 2006. Neighborhood design and vehicle type choice: evidence from northern California. *Transportation Research Part D: Transport and Environment*, 11(2), pp.133–145.

Dubin, R.A., 1998. Spatial autocorrelation: a primer. *Journal of Housing Economics*, 7(4), pp.304–327.

Feld, S.L., 2010. Why Your Friends Have More Friends than You Do1. , 96(6), pp.1464–1477.

Gilbert, E. & Karahalios, K., 2009. Predicting tie strength with social media. *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, p.211.

He, Y., Zhang, C. & Ji, Y., 2012. Principle Features for Tie Strength Estimation in Micro-blog Social Network. *2012 IEEE 12th International Conference on Computer and Information Technology*, pp.359–367.

Kim, J. & Goldsmith, P., 2008. A spatial hedonic approach to assess the impact of swine production on residential property values. *Environmental and Resource Economics*, 42(4), pp.509–534.

LeSage, J.P. & Pace, R.K., 2009. *Introduction to Spatial Econometrics*, Chapman and Hall/CRC.

Pinjari, A.R. et al., 2011. Modeling the choice continuum: an integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions. *Transportation*, 38(6), pp.933–958.

Tobler, W.R., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46, pp.234–240.

Walker, J. & Ben-Akiva, M., 2002. Generalized random utility model. *Mathematical Social Sciences*, 43, pp.303–343.

Due, P., et al. (1999). "Social relations: network, support and relational strain." Social Science and Medicine **48**: 661-677.

Facebook (2015). "Company Info." Newsroom. Retrieved 1/18/2015, 2015, from http://newsroom.fb.com/company-info/.

Lewis, K., J., et al. (2008). "Tastes, ties, and time: A new social network dataset using Facebook.com." Social Networks **30**(4): 330-342.

O'Reilly, P. (1988). "Methodological issues in social support and social network research." Social Science and Medicine **26**: 863-873.

Smith, A. (2011). Why Americans use social media. Pew Internet and American Life Project. Washington, DC, Pew Research Center.

Twitter (2015). "About." Retrieved 1/18/2015, 2015, from https://about.twitter.com/company.

Wall, T. A., et al. (2014). "Exploring the Use of Egocentric Online Social Network Data to Characterize Individual Air Travel Behavior." Transportation Research Record **2400**: 78-86.