Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke

The Role of Geography in CouchSurfing Social Network Activities

Edward Pultar
Institute of New Imaging Technologies
Universitat Jaume I
Castellón, Spain
pultar@uji.es

Martin Raubal
Institute of Cartography and Geoinformation
Swiss Federal Institute of Technology, ETH
Zürich, Switzerland
mraubal@ethz.ch

Konstadinos Goulias
Department of Geography
University of California, Santa Barbara
California, USA
goulias@geog.ucsb.edu

Abstract

Social networks are ubiquitous in today's mobile information society. This research focuses on social networks that depend on the physical and virtual locations of their users in order to provide various services. In these contemporary social networks both virtual and physical presence is a requirement. This research examines travel behavior using an Internet-based website, CouchSurfing, which provides free lodging with local residents. Increases in computing power and accessibility have led to novel e-travel techniques and the users of such systems utilize an amalgamation of social networks, transportation networks, and data communication networks. Thus the focus is on how the geographical spread of people in a modern, digital social network influences the travel choices of each individual in the network. Data was collected from participants in more than 35 different countries to examine the effects of geography and demographics on social network travel behavior. This research emphasizes novel movement techniques utilizing Internet-based social networking which leads to new conclusions about the power and potential for contemporary social networks.
Keywords : social networks, geography, travel behavior.

1 Introduction and Background

Social networks have a key influence on modern travel behavior. There is a contemporary need for understanding the effects of social connections on travel activities [1]. This research addresses the need for studying international travel behavior utilizing social networks.

A collection of people is required to exhibit the role of a geographically distributed social network on travel behavior. Here, we utilize an existing social network, CouchSurfing (CS, http://www.couchsurfing.org). The CS network has over 3 million members as of 2011 and serves as a resource for hospitality exchange around the globe. Those in the CS network permit others to stay at their residence. This may be in the form of a couch, sofa, futon, or an extra bed. Thus members perform two distinct activities: 1) being a guest and 2) being a host. These hospitality exchanges currently take place in over 200 different countries.

Initially a guest contacts a host by sending a message through the social network system. A face-to-face meeting is scheduled if it is permitted by the spatio-temporal constraints of the guest and host. After the guest leaves the home of the host each individual writes a review (positive or negative) of the other using the social network. Hence the activities in this travel behavior move between the virtual and physical worlds. The opening contact is virtual and then the travel activity moves to the physical world when the guest and host meet. Following the meeting and hosting there is another virtual contact component that occurs via reference writing.

Each individual resides in a specific location and a potential guest may specify geographic or other search criteria, such as age and gender, for potential hosts. There is additional information on each member's profile about the individual such as: favorite activities, life philosophies, pictures, and references left by other members. After a face-to-face meeting in a physical location the social networking moves back to the virtual world as reviews are left for one another that will affect one's reputation and ability to use the service in the future. Members often start using the network by either exclusively being a guest or hosting others but later engage in both as is necessary for the network to succeed [2]. While members initially interact through virtual means, the core purpose of the network is an affordance of physical interaction, blending physical and virtual social networks [3].

Every network member creates a profile containing Volunteered Geographic Information (VGI) [4] such as their current city of residence, previous travels, and future travel plans. A profile can also contain personal descriptions, hosting capabilities, languages spoken, and photographs. The CouchSurfing social network is increasingly important as the links created between nodes can be used to gather a level of safety for travelers as well as a global map with the geographic locations and places of connected friends. Members began signing up for the CS project in 2004 with hundreds joining per month. By 2005 thousands were joining each month and steady growth has continued over the years right up to 2011 when over ten thousand people were joining per month. Now with over 3 million members from all around the planet this social network continues to influence travel behavior by modifying perceptions and abilities for individuals around the earth.

Users of CS now cover a vast portion of the globe with members on each continent. It is a contemporary, emerging form of connection. Further expansion of this unique traveler system is expected to carry on well into the future; therefore any results discovered in this research afford a utility for future developments. This utility may be in the form of understanding different roles individuals play in a social network and what characteristics play a part in interactions.

The next section discusses count data models and Section 3 presents the data collection process. The fourth section contains the count-based model analysis and the last section finishes with conclusions and future work.

2 Count Data Models

This section uses survey methods to further examine how characteristics of CouchSurfers affect their use of the network.

The work was done to examine real world members of the social network and their different levels of involvement. A survey was performed that gathered data from over 200 participants. Hence the sample data is drawn from the population of CouchSurfing (CS) network members. These data were then used with count-based models to see effects of demographic and geographic variables on CS activities.

Statistical methods are used in a large assortment of cases to identify, study, and solve a variety of problems in transportation [5]. Data are collected and analyzed to aid in decision-making and understand complex processes. There are also future opportunities for social network researchers to utilize statistical methods for analyzing social network issues such as network measures and structures [6].

The notion of count data models stems from the need to model processes that involve counts of activities such as trips. Hence these counts are positive integers describing something such as the number of times an individual has participated in an activity. Therefore, using standard regression techniques is inappropriate, as it would yield incorrect results by predicting negative values for scenarios involving positive counts [7].

The Poisson regression model from [7] is defined as:

$$P(y_i) = \frac{EXP(-\lambda)\lambda_i^{y_i}}{y_i!} \qquad (1)$$

with $P(y_i)$ being the probability of having y_i events. For example, if one were researching grocery store robberies $P(y_i)$ would be the probability that grocery store i would have y_i robberies per year. Poisson parameter λ_i is the expected amount of robberies per year at grocery store i where

$$\lambda_i = EXP(\boldsymbol{\beta}\mathbf{X}_i) \qquad (2)$$

and

$$LN(\lambda_i) = \boldsymbol{\beta}\mathbf{X}_i \qquad (3)$$

with $\boldsymbol{\beta}$ being the vector of coefficients and $\mathbf{X}_i$ the explanatory parameters. To estimate this model standard maximum likelihood methods are used with likelihood:

$$L(\boldsymbol{\beta}) = \prod_i \frac{EXP\left[-EXP(\boldsymbol{\beta}\mathbf{X}_i)\right]\left[EXP(\boldsymbol{\beta}\mathbf{X}_i)\right]^{y_i}}{y_i!} \qquad (4)$$

and log likelihood function:

$$LL(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[-EXP(\boldsymbol{\beta}\mathbf{X}_i) + y_i\boldsymbol{\beta}\mathbf{X}_i - LN(y_i!)\right] \qquad (5)$$

Ideally with the Poisson model the mean and variance are to be equal: E[yi] = VAR[yi]. However, if the mean is greater than the variance the data is underdispersed and if the mean is less than the variance the data is overdispersed. Examples and discussion of dispersion can be found in [8]. The negative binomial model addresses this issue and is formed by changing Equation 2 to look as follows:

$$\lambda_i = EXP(\boldsymbol{\beta}\mathbf{X}_i + \varepsilon_i) \qquad (6)$$

with the new term being gamma-distributed with mean 1 and variance α^2. This allows the mean and variance to differ resulting in:

$$VAR[y_i] = E[y_i][1 + \alpha E[y_i]] = E[y_i] + \alpha E[y_i]^2 \qquad (7)$$

Hence there are different methods available within the family of count-based models and the researcher must make an informed decision as to which is most appropriate, for an example see [9]. This decision is based on issues such as the knowledge of the activity behavior, the dataset, influences of each variable, and feedback from testing a variety of methods.

3 Data Collection

In this work 228 CouchSurfing (CS) members from over 35 countries (distribution discussed later with Table 2) completed a survey to address the following hypotheses:

- Members of the CS social network with more years of age tend to host more than younger members.
- Male users tend to play the role of guest more often than females.

Hence the hypotheses state that hosts are more commonly older and female while guests are more likely male and have less years of age. The motivation for these hypotheses comes from the desire to understand what individual characteristics play a role on a person's travel behaviour. This is also important for seeing how well the network is functioning and if some opportunities are more available to certain individuals. This survey's questions pertained to the number of trips a member had taken in the past year using CouchSurfing, i.e., the count of how many times they had been a guest. The participants also provided the count of how many times they had been a host in the past year.

Demographic data were also collected:

- Age
- Gender
- Current country of residence

Before launching the survey publicly a collection of Santa Barbara, California, USA, CouchSurfers was selected as a research focus group. These ten individuals were then invited to appear in person for the focus group gathering where they received a free meal and the survey was further examined and discussed. After this meeting, the survey was posted in message forums on the CouchSurfing.org website with no incentives and resulted in 228 participants.

The 228 observations in this dataset had an average age of 32.3 and 38% of the participants were male meaning 62% were female. Females in this collected dataset were more willing to give their time since they represented 43% of the 2.5+ million members of CouchSurfing on 1 March, 2011. In future surveys, techniques may be used to ensure the dataset has more males than females. The ages of individuals that completed the survey varied between 19 and 65 whereas the ages of the 2.5+ million members of CS range between 18 and 89. Last, the average age in this dataset, 32, was slightly higher than 28, which is the average age of the 2.5+ million CouchSurfers in March 2011. The full descriptive statistics are in Table 1 and the variables are described below:

- GUEST: Number of times the individual has been a guest in previous 365 days
- HOST: Number of times the individual has been a host in previous 365 days
- AGE: Individual's age

Table 1: Descriptive statistics for sample containing 228 observations.

Variable	Mean	Std. Dev.	Minimum	Maximum	Cases
GUEST	7.87	10.377	0	60	228
HOST	12.29	18.379	0	100	228
AGE	32.30	11.487	19	65	228
MALE	.38	.487	0	1	228

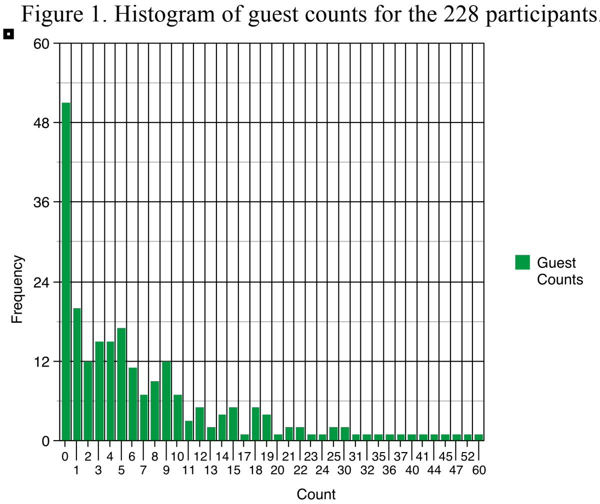
- MALE: 1 for male, 0 for female

4 Analysis with count-based models

Count-based models for this analysis were implemented in LIMDEP's software package NLOGIT 4.0.1. Demographic and place of residence variables are used to describe guest and host activities. This was done using Poisson and Negative Binomial models. This brings about the concept of Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) models. These Zero-Inflated models handle situations where zero values of counts (guest and host activities here) can occur from qualitatively different situations [7]. One situation for a zero value in the host count is that the individual has not hosted during the observation period (previous 365 days). A different situation for a zero value is the lack of an ability to host due to conditions such as unwelcoming roommates. Hence zero-inflated models will be examined with the collected dataset.

The histogram in Figure 1 shows the distribution of the data for guest events. Here it is apparent there is inflation in zero values, e.g., 51 of the 228 individuals in the study had not participated in a guest event in the past year. This may be for qualitatively different reasons as explained in the previous paragraph but these graphs further reiterate potential use of zero inflated methods of analysis.

Figure 1. Histogram of guest counts for the 228 participants.



The following analysis uses explicit geographic details derived from the country of residence variable collected for each of the 228 survey participants. The countries of residence for the participants are listed in Table 2 along with a count of how many are from each place. Table 3 shows the counts of participants per continent.

Table 2. Counts of survey participants by residence country.

Country	Count	Percent of Sample
Argentina	4	1.7
Australia	1	.4
Austria	19	8.3
Brazil	20	8.7
Canada	28	12.2
Cape Verde	1	.4
China	2	.8
Colombia	4	1.7
Costa Rica	1	.4
Denmark	1	.4
France	10	4.3
Germany	17	7.4
Ghana	1	.4
India	4	1.7
Indonesia	9	3.9
Ireland	2	.8
Israel	16	7.0
Italy	12	5.2
Moldova	1	.4
Morocco	1	.4
Netherlands	3	1.3
New Zealand	1	.4
Oman	1	.4
Peru	1	.4
Philippines	1	.4
Poland	2	.8
Romania	1	.4
Russia	1	.4
Serbia	1	.4
Singapore	7	3.0
Slovakia	2	.8
South Africa	5	2.1
Spain	1	.4
Sudan	1	.4
Sweden	1	.4
Switzerland	1	.4
Turkey	4	1.7
United Kingdom	7	3.0
USA	30	13.1
Vietnam	3	1.3

Table 3. Survey participants with continent of residence.

Continent	Count	Percent of Sample
Africa	9	3.9
Asia	48	21.0
Europe	81	35.5
Latin America	30	13.1
North America	58	25.4
Oceania	2	.8

The following six countries were chosen for further analysis due to their counts being 15 or greater: Austria, Brazil, Canada, Germany, Israel, USA. Binary dummy variables were created for each of the respective countries and count-based models were used to examine whether or not these countries of residence play a role in guest and host activity counts. Here the Zero-Inflated Poisson model was used as it is appropriate given the Vuong statistic (used to test whether or not to use a zero-inflated model, [10]), significance of variables, and predicted number of zero counts. In Figures 2 and 3 are the ZIP results for guest and host counts using the country variables along with those of age and gender.

Figure 2: ZIP for Guest Counts with Country of Residence.

Zero Altered Poisson Regression Model

Comparison of estimated models

	Pr[0|means]	Number of zeros	Log-likelihood
Poisson	.001	Act.=51 Prd.=.1	-1497.333
Z.I.Poisson	.223	Act.=51 Prd.=50.9	-1188.873

Vuong statistic for testing ZIP vs. unaltered model is 5.599.
Distributed as standard normal. A value greater than +1.96 favors the zero altered Z.I.Poisson model. A value less than -1.96 rejects the ZIP model.

Poisson/NB/Gamma regression model

Variables	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
Constant	2.15	.028	77.359	.000	
MALE	.34	.016	20.990	.000	.38
AGE	-.00	.001	-.253	.800	32.30
AUSTRIA	.24	.021	11.357	.000	.08
BRAZIL	.00	.034	.025	.980	.09
CANADA	-.01	.028	-.370	.711	.12
GERMANY	.30	.024	12.166	.000	.07
ISRAEL	-.06	.041	-1.387	.166	.07
USA	-.17	.033	-5.282	.000	.13

Figure 3: ZIP for Host Counts with Country of Residence.

Zero Altered Poisson Regression Model

Comparison of estimated models

	Pr[0|means]	Number of zeros	Log-likelihood
Poisson	.000	Act.=52 Prd.=.0	-2508.460
Z.I.Poisson	.223	Act.=52 Prd.=52.1	-1928.789

Vuong statistic for testing ZIP vs. unaltered model is 5.923.
Distributed as standard normal. A value greater than +1.96 favors the zero altered Z.I.Poisson model. A value less than -1.96 rejects the ZIP model.

Poisson/NB/Gamma regression model

Variables	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
Constant	2.23	.017	131.590	.000	
MALE	.19	.001	22.948	.000	.38
AGE	.01	.000	30.998	.000	32.30
AUSTRIA	-.08	.018	-4.669	.000	.08
BRAZIL	.07	.014	5.679	.000	.09
CANADA	-.20	.016	-12.632	.000	.12
GERMANY	.44	.012	37.445	.000	.07
ISRAEL	-.47	.039	-12.234	.000	.07
USA	-.21	.016	-13.573	.000	.13

The guest model has a positive gender coefficient as does the host model meaning guests and hosts are more likely to be males. However, the age variable is not significant at the 5% level in the guest model here hence no conclusions are made with respect to age. The country variables significant in both guest and host models are Austria, Germany, and USA. Between these three countries the coefficient for the USA is the lowest, suggesting that in this sample the social network members from the USA are participating in less guest and host activities compared to those in Austria and Germany as well as the countries not included in the model which have a coefficient of zero. This may be due to contemporary travel trends or the recent status of economies. Also of note here is how much German members stand out with respect to hosting activities, a much higher coefficient than Austria and the USA, suggesting Germany is quite a hospitable region. With the sample size of 228 the counts of individuals in these countries of residence were all 30 or less (Table 2) and this resulted in a number of variables not being significant. The next models use the continent of residence data shown in Table 3, which have larger counts of individuals.

These four continents from Table 3 were chosen for further investigation in a count-based model scenario: Asia, Europe Latin American, and North America. Each of the four continents has its corresponding binary dummy variable that is used in the ZIP model along with age and gender variables. The predicted number of zero counts, Vuong statistic, and significance of variables suggest a fit for the ZIP model as shown in Figures 4 and 5.

models using continent data and the change of sign (-.002 and .015) between guest and host indicates that the older a person gets the less they will participate in guest activities but they will participate in more hosting activities. Some explanations for this may be due to mobility at different ages and/or stages of life. Additionally, an older person has a higher likelihood of having accumulated more resources to share with others. The continent data shows Europeans being the most active in this sample for both guest and host activities whereas the least active here are those in North America. Another interesting point is where Latin America stands in its performing of guest vs. host activities. In the guest model the coefficients for the

Figure 4. ZIP for Guest Counts with Continent of Residence

Zero Altered Poisson Regression Model

Comparison of estimated models

	Pr[0|means]	Number of zeros	Log-likelihood
Poisson	.001	Act.=51 Prd.=.1	-1460.341
Z.I.Poisson	.218	Act.=51 Prd.=49.8	-1167.292

Vuong statistic for testing ZIP vs. unaltered model is 5.861.
Distributed as standard normal. A value greater than +1.96 favors the zero altered Z.I.Poisson model. A value less than -1.96 rejects the ZIP model.

Poisson/NB/Gamma regression model

Variables	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
Constant	1.47	.097	15.171	.000	
MALE	.34	.015	22.798	.000	.38
AGE	-.00	.001	-3.040	.002	32.30
ASIA	.67	.094	7.165	.000	.21
EURO	.97	.095	10.216	.000	.36
LATAMER	.68	.097	6.928	.000	.13
NTHAMER	.63	.093	6.775	.000	.25

Figure 5. ZIP for Host Counts with Continent of Residence

Zero Altered Poisson Regression Model

Comparison of estimated models

	Pr[0|means]	Number of zeros	Log-likelihood
Poisson	.000	Act.=52 Prd.=.0	-2467.595
Z.I.Poisson	.228	Act.=52 Prd.=52.0	-1903.457

Vuong statistic for testing ZIP vs. unaltered model is 6.121.
Distributed as standard normal. A value greater than +1.96 favors the zero altered Z.I.Poisson model. A value less than -1.96 rejects the ZIP model.

Poisson/NB/Gamma regression model

Variables	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
Constant	1.13	.030	37.496	.000	
MALE	.20	.008	23.165	.000	.38
AGE	.02	.000	-39.784	.000	32.30
ASIA	.97	.023	41.407	.000	.21
EURO	1.25	.023	53.762	.000	.36
LATAMER	1.06	.026	40.974	.000	.13
NTHAMER	.81	.023	35.545	.000	.25

The modeling of Figures 4 and 5 agrees with the previous since in this sample guests and hosts tend to be male more than female. The age variable is significant in both ZIP

continents of Asia (.671), Latin America (.675), and North America (.633) do not vary greatly but in the host model Latin America separates itself more from North America and Asia (1.058 for Latin America vs. .813 for North America and .968

Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke

for Asia) suggesting that Latin America engages in more hosting activities than the other two continents. This hints at members of the CouchSurfing social network in Latin America being more hospitable but also may be a result of economic restraints on traveling abroad. For example, countries in Latin America are not members of the USA visa waiver program. Hence if members in these locations wish to participate in this social network it may be easier for them to play the role of host vs. guest. Last, this model suggests that Asia, Europe, Latin America, and North America with their positive coefficients have higher frequencies of guest and host counts than Africa and Oceania.

5 Conclusions and Future Work

Count-based models have been used with the gathered sample to examine characteristics that affect how much an individual performs host and guest activities. This is important for making future changers to the network and travel policies in order to equally distribute possibilities for people around the world. Thus this analysis may have an impact on better catering to people of all geographies, genders, and ages to allow more international exchange opportunities. In future research more variables (such as positive, negative experiences, strength of connections) will be collected to further determine influences for an individual to perform one role in a social network vs. a different role.

Regarding the initial hypotheses, the count-based models have suggested that in this sample older members are hosting more and male users are being guests and hosts more often than females. Future work may also be done to examine as to why males are traveling and hosting more than females. It is also of note that the estimates the models provide are indeed estimates and more work with random, larger samples will be helpful in further investigating the hypotheses and geographies.

Some ideas for why these differences in age and gender are happening are given above but a deeper explanation is valuable to further fine-tune the social network as a whole such that it continues to be successful. Even though CouchSurfing members are the greater population studied here these results may be of use to other travel and social networks. The results can help individuals associated in all facets of tourism (such as transport, lodging, alimentation, tours) to better understand how to serve their clients and specifically what are the characteristics of the clients being served for using with tools such as a dynamic GIS [11]. Future samples can be gathered from other populations in order to explore the differences and similarities in various populations.

The use of geographic information in the form of countries and continents suggests how the location where a person lives can affect their role and activity in a global, Internet-based social network. All of this comes together and highlights the intrinsic connection between virtual and physical presence in modern, Internet-based social networks.

References

[1] CARRASCO, J. A., MILLER, E. J. & WELLMAN, B. 2006. Spatial and Social Networks: The Case of Travel for Social Activities. 11th International Conference on Travel Behaviour Research. Kyoto, Japan.

[2] LAUTERBACH, D., TRUONG, H., SHAH, T. & ADAMIC, L. 2009. Surfing a web of trust: Reputation and Reciprocity on CouchSurfing.com. IEEE Social Computing 2009. Vancouver, British Columbia, Canada.

[3] PULTAR, E. & RAUBAL, M. 2009. A Case for Space: Physical and Virtual Location Requirements in the CouchSurfing Social Network. In: XIE, X. Z. A. X., ed. International Workshop on Location Based Social Networks (LBSN'09) at 17th ACM International Symposium on GIS, SIGSPATIAL ACM GIS 2009, Seattle, Washington, USA, 88-91.

[4] GOODCHILD, M. F. 2007. Citizens as sensors: the world of volunteered geography. GeoJournal, 69, 211-221.

[5] GOULIAS, K. G. 2003. Transportation systems planning: methods and applications, Boca Raton, Fla., CRC Press.

[6] KNOKE, D. & YANG, S. 2008. Social Network Analysis, California, Sage Publications, Inc.

[7] WASHINGTON, S., KARLAFTIS, M. & MANNERING, F. 2003. Statistical and Econometric Methods for Transportation Data Analysis, Washington, DC, Chapman & Hall.

[8] KARLAFTIS, M. G. & TARKO, A. P. 1998. Heterogeneity considerations in accident modeling. Accident Analysis and Prevention, 30, 425-433.

[9] MA, J. & GOULIAS, K. G. 1999. Application of poisson regression models to activity frequency analysis and prediction. Transportation Research Record 1676, 86-94.

[10] VUONG, Q. 1989. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica, 57, 307-334.

[11] PULTAR, E.F., COVA, T.J., YUAN, M., & GOODCHILD, M.F. 2010. EDGIS: A Dynamic GIS Using Space Time Points. International Journal of Geographical Information Science, 24, 329-346.