Implementing an Agent-Based Formation of Social Networks for Joint Travel

Gabriel Hannon*[1], Joanna Ji[2], Qin Zhang[3], Dr. Ana Tsui Moreno[4], and Dr. Rolf Moeckel[5]

[1,2,3,4,5]School of Engineering and Design, Technical University of Munich, Germany

Short summary

Introducing social networks to travel demand models could better capture socially induced travel behavior. This paper presents an agent-based approach to forming social networks that match important global characteristics and egocentric homophilies in distance, age, and gender for a population on the order of 10^6. Based on data from an egocentric snowball sample, this methodology successfully reproduces homophilies in age and gender, as well as an expected power-law distribution of geographic distance between connections. An initial clique formation heuristic is implemented on top of the homophily calculations. The generated network exhibits preferential attachment between agents of higher degree, in line with more general literature on network formation.

Keywords: agent-based modeling, synthetic social networks, synthetic populations, transportation network modeling, travel demand modeling

1 Introduction

As Frei and Axhausen put it in their seminal 2007 paper on the geography of social networks, "travel is the price we pay to be with others" (Frei & Axhausen, 2007). Recently, researchers in transportation modeling have further investigated the marked interactions between social connections and travel activity participation (Carrasco et al., 2008; Kim et al., 2018). Social contacts affect multiple aspects of travel behavior, from activity generation to destination and mode choice (Kim et al., 2018). The literature has identified socio-demographic homophilies, geographic distance, and structural network properties, such as clique and degree distributions, as key characteristics of travel-focused social networks (Arentze et al., 2012; Illenberger et al., 2013; Dubernet, 2017). The synthesis of a static social network for a study area is a key next input for generating joint travel demand.

Social network structures are studied across fields. The development of a synthetic social network suitable for travel behavior modeling, however, has been challenging. Illenberger et al. (2013) approach this problem with an exponential random graph model, which accounts for homophily but has limited transitivity. Arentze et al. (2012) explicitly account for transitivity via a link probability model using binary logit estimations for forming friendships. Work by Dubernet (2017) generates a social network with a heuristic that accounts for socio-demographic and geographic homophily in addition to clique distribution, but does not account for agent-specific preferences. All three of the above utilize the snowball data sampled by Kowald & Axhausen (2012).

Building on this research, we attempt a scalable method to generate a synthetic social network that matches known socio-demographic homophily, connection distance, reciprocity, and transitivity properties. As the network is eventually intended to couple with a travel demand generation model, it is synthesized explicitly for the Munich metropolitan region.

2 Methodology

The fundamental methodology relies on a union of agent-based objects and network data structures. Agent objects are constructed from a synthetic population for the city of Munich generated by the open-source land-use model SILO (Moreno & Moeckel, 2018). Network and agent attributes - namely degree distribution, edge length preferences, and egocentric homophilies - are derived from ETH-Zurich's snowball dataset (Kowald & Axhausen, 2012). This section describes these input

data and the social network formation algorithm.

Input data

Our study area is the Munich metropolitan region, including the City of Munich and the surrounding cities of Augsburg, Ingolstadt, Landshut, and Rosenheim, as seen in Figure 1. This area has a population of 4.5 million in roughly 2.1 million households. We use the output of the SILO land use model, which generates a synthetic population for the region based on census data and Iterative Proportional Updating (Moreno & Moeckel, 2018). Though a level of social connection can be inferred from shared households, workplaces and education places, there are no additional social or friendship connections. Therefore, we set out to generate a friendship network for a 5% sample of this synthetic population.

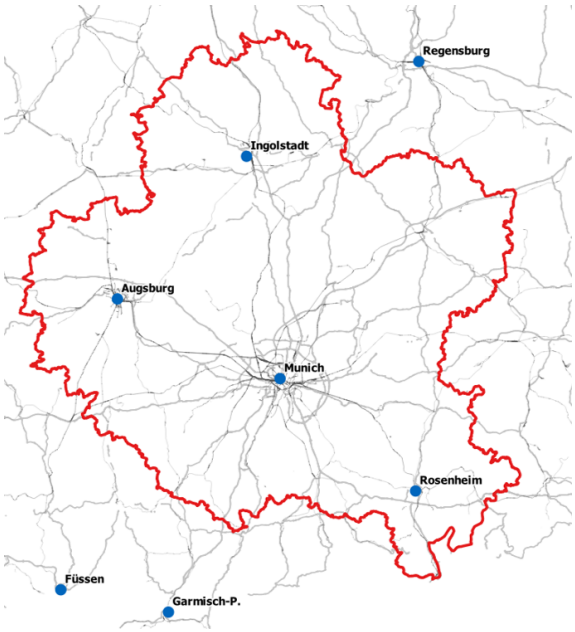

Figure 1: Munich metropolitan region.

Snowball Data

Despite the proliferation of data regarding large social networks, only a few datasets include the type of geographically-embedded demographic connection data that may impact joint leisure travel. Kowald & Axhausen's (2012) snowball survey provides a key source for this topic. After data cleaning, this survey yields a total of 793 egos and 14,326 unique 'names.'

While this data is extensive, it maps a sparse network, which tends to branch out to isolated alters. Therefore, any analysis of this data must focus on egocentric metrics, not global network characteristics. Previous work has established that age, gender, and distance homophilies are the most significant demographic attributes in leisure travel networks (Dubernet, 2017; Arentze et al., 2012; Illenberger et al., 2013). These characteristics, along with ego degree, form the main egocentric attributes for our formation model.

While previous work has used population-aggregated homophilies (Kowald & Axhausen, 2012; Dubernet, 2017), further examination suggests that demographic attributes affect an agent's willingness to accept variation across these homophilies. After dividing the data into segments based on eight 10-year age brackets and two genders, Figure 2 illustrates the distribution in accepted average gender homophily, age difference, and degree by segment. Such segmenting shows, for example, that egos aged over 70 ($70M$, $70F$, $80M$ or $80F$) tend to accept ties across greater ranges of age differences and have smaller degrees. These insights motivate a formation model that accounts for the interaction between an agent's demographic characteristics and homophily profile.

This segmenting leads to distributions for agent degrees by segment and a table of distributions for age and gender homophilies by segment pair. Figure 3 compiles the age and gender homophily distribution table. Rows represent network-level distributions of connections from egos in one

Figure 2: Segmented distribution in age difference, gender homophily, and degree for snowball egos by segment. A gender homophily of 1 indicates similar genders.

segment to alters in all other segments. This matrix is notably asymmetric; we can attribute this to sampling bias in the snowball data - e.g., 28% of egos are in their 50s while only 0.6% are younger than 20 - and fundamental asymmetries in the way that connections and popularity are distributed among agents in all social networks (Barabási & Albert, 1999).

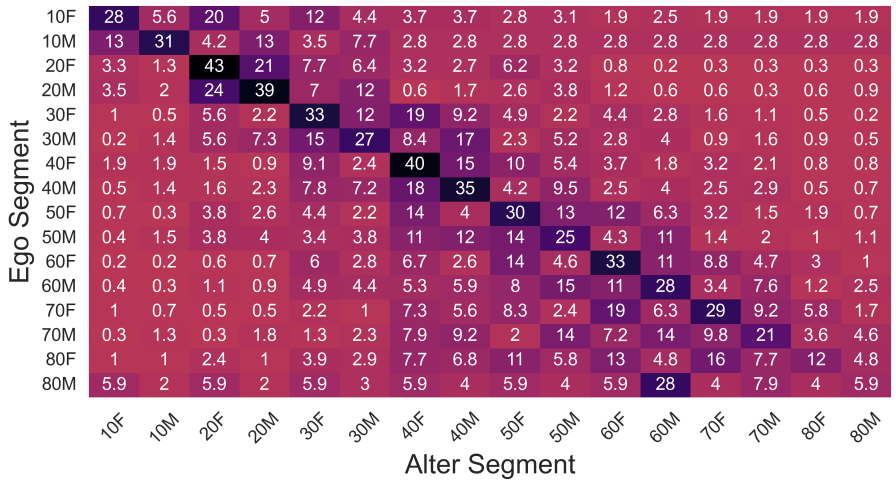

Figure 3: Segment to segment homophily distributions. Rows egocentrically represent the desired distribution by percentage. Only rows sum to 100%.

Distance data, however, is addressed at the network-level through a power-law probability distribution function for all edges, as initially demonstrated in Illenberger et al. (2013). The probability of forming an edge of distance d between any two agents is described by the following distribution function:

$$P(\text{Edge Formation}|d) = \frac{(\alpha - 1)x_0^{(\alpha-1)}}{d^{\alpha}} \tag{1}$$

We specify the exponential and scale factor parameters, α and x_0, by a least squares estimation on the distribution of snowball distance edges that would fit inside the Munich study area. An

additional normalizing factor is introduced in the code to ensure that the cumulative distribution function converges to one within the study area.

Social network formation

The social network generation algorithm uses the segmented snowball data distributions and the SILO synthetic population as inputs. The formation method broadly follows the iterative three-step process displayed in Figure 4. In short, a friend-goal dictionary is generated for each agent based on the agent's segment-based age and gender homophilies. Then, the algorithm matches two mutually compatible agents based on these dictionaries. A connection is only formed if the agents pass a stochastic draw based on the geographic distance between them. If a connection forms, each agent updates their friend-goal dictionary before searching for further connections. After all agents have entered the matching stage or no new connections can be formed, unsatisfied agents redraw their friend-goal dictionaries, restarting the cycle.

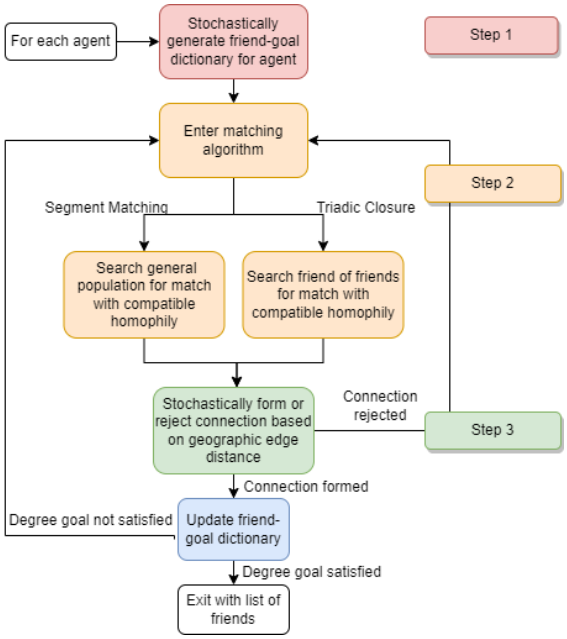

Figure 4: Social network generation steps for each iteration. Friend goal dictionary periodically re-drawn to avoid mismatches.

More specifically, in Step 1, each agent calculates how many degrees, or friends, k, they need to reach their total degree goal, which is assigned during population synthesis. Each agent then draws k times from their homophily distribution to form their 'friend-goal' dictionary. This provides the segments from which each agent is willing to search for connections during Step 2. Figure 5 illustrates two example agents after an initial draw.

Agent A	**Gender**		**Age**		**Segment**										**Current degree**
	Female		41		40F		*eligible segments highlighted								**0**
Segments eligible for connection	**10F**	*10M*	**20F**	*20M*	*30F*	**30M**	**40F**	**40M**	**50F**	*50M*	**60F**	*60M*	**70F**	*70M*	*80F*	**Degree goal**
Number of desired connection	**1**	*0*	**1**	*0*	*0*	**2**	**8**	**1**	**2**	*0*	**1**	*0*	**1**	*0*	*0*	**17**
Agent B	**Gender**		**Age**		**Segment**										**Current degree**
	Female		22		20F										**0**
Segments eligible for connection	*10F*	*10M*	**20F**	**20M**	**30F**	**30M**	**40F**	*40M*	**50F**	*50M*	*60F*	*60M*	*70F*	*70M*	*80F*	**Degree goal**
Number of desired connection	*0*	*0*	**3**	**2**	**1**	**2**	**1**	*0*	**1**	*0*	*0*	*0*	*0*	*0*	*0*	**10**

Figure 5: Sample agents after initial draw.

In Step 2, the algorithm matches mutually eligible agents. For example, in Figure 5, Agent A is in the $40F$ segment and is looking for one connection in the $20F$ segment; Agent B is in the $20F$ segment and is looking for a connection in the $40F$ segment. They are mutually compatible and could be matched.

During Step 2, agents search for matches from the general population *or* from their 2^{nd}-degree connections - i.e., friends of their friends. The latter acts as a heuristic for triadic closure. Any con-

nection formed by triadic closure automatically creates a clique of at least three while maintaining the proper homophilies for all clique members; this provides an agent-based implementation of a concept explored in Asikainen et al. (2020).

The algorithm currently forms an initial proportion of agents' degree goals by matching from the general population. It then switches to triadic closure, which reduces the search space but enables clique formation. When formation progress stagnates, the algorithm switches back to the general population search to complete the remaining connections.

After any potential match is found, this connection enters Step 3 where it is accepted or rejected based on the distance between the two agents. Equation (1) generates a connection probability for this distance, which is then compared to the result of a uniform draw from $(0, 1)$. If the edge is formed, each agent removes one from the appropriate segment of their friend-goal. If this completely satisfies either agent's degree goal, that agent exits the algorithm. Otherwise, each agent returns to Step 2 to search for new connections.

After an iteration exhausts its possible matches, all remaining agents stochiastically redraw their friend-goal dictionaries via Step 1 and the process repeats. Because of asymmetries in the homophily matrix and the demographic variation between the snowball and synthetic populations, these stochastic iterations are necessary for everyone to achieve their desired degree. By repeatedly drawing from the segment homophilies, the algorithm induces an equilibrium that balances these asymmetries.

3 Results and discussion

The current algorithm is built in Python and makes use of the Networkx package (Hagberg et al., 2008). The algorithm scales roughly as $\mathcal{O}(n^2)$ and runs in 30 minutes, generating $1,786,885$ edges for $202,401$ nodes.

Segment Based Homophilies

The aggregate segment-level connections are demonstrated in Figure 6. In general, the core adult population homophily distributions are well obeyed; this makes sense as they are well represented in the snowball data and have sufficient connection possibilities within their desired segments. The significantly older and significantly younger segments, which were underrepresented in the input data, perform worse. For example, segment $10F$ has too many internal connections and not enough connections to older segments; simultaneously, the other segments largely meet their goals regarding segment $10F$. This ties back to the asymmetries in the homophily matrix (Figure 3) where, for example, $M_{10F,30F} = 12\%$ but $M_{30F,10F} = 1\%$. In these cases, one segment meets its goal easily and has no need to surpass it, while the other tends to form more internal connections.

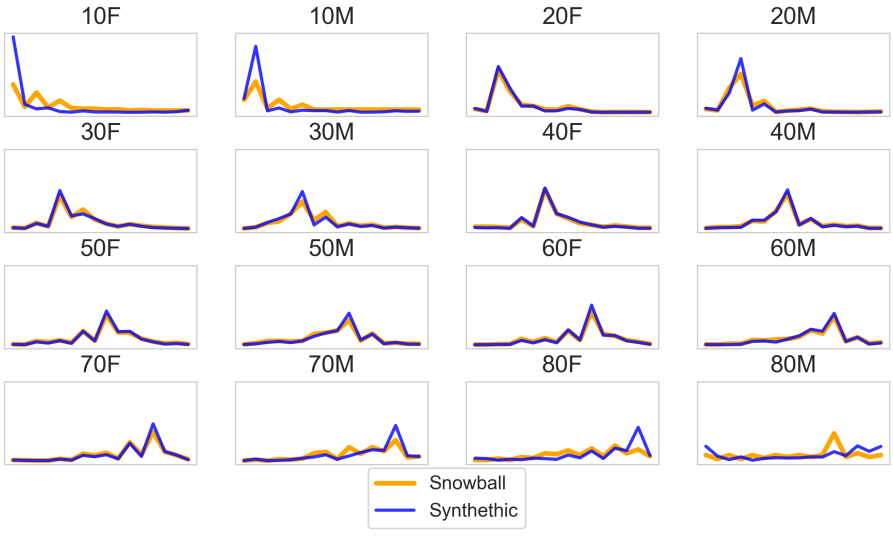

Figure 6: Input and realized homophily distributions by segment.

Figure 7 summarizes the egocentric distribution of gender similarity and age difference. While the synthetic network does exhibit slightly higher degrees of homophily - i.e., higher gender similarity and lower average age difference - it still captures activity at the tails of each distribution.

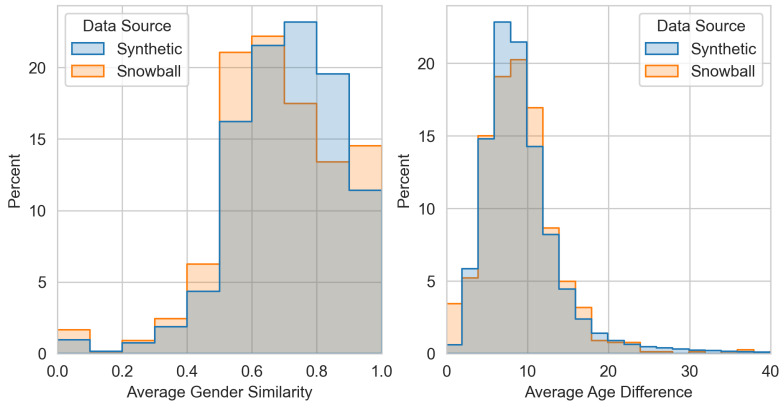

Figure 7: Average egocentric gender and age homophilies.

Distance Distribution Matching

Overall, the approach to distance formation reliably replicates the expected geographic edge length distribution. Figure 8 demonstrates the realized distribution of edge lengths for the entire network compared to the snowball data. There is a slight under-representation of edges less than 3 km as the basic power law formulation cannot be sensitive to distances smaller than its scale parameter, x_0 - which was set as ~ 1.5 km based on the snowball data - without causing its probability integral to diverge. Additionally, the synthetic distribution approaches 100% earlier than the snowball data, though this is likely an edge effect, given that only agents near the border of the study area can form the longest connections.

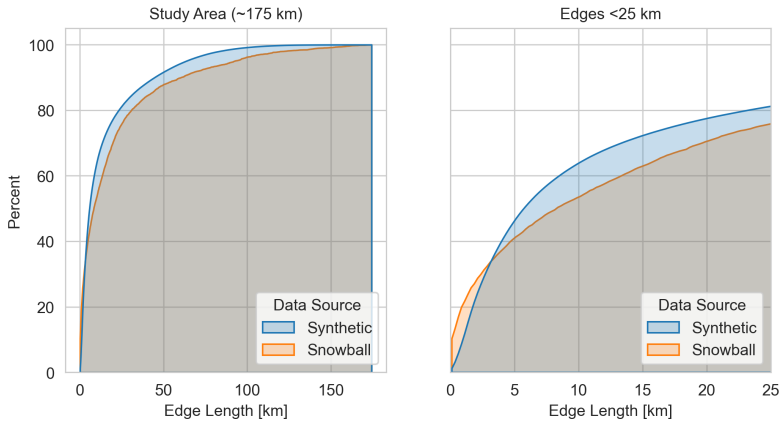

Figure 8: Edge length distribution for synthetic population.

Initial Clique Formation

The initial triadic closure methodology provides a basic mechanism for clique formation. Roughly 82% of agents have at least one clique in the full network, though the method significantly overforms small cliques compared to the input data. While 2^{nd}-degree connections who have multiple mutual friends with the searching agent are duplicated in the triadic closure search space, providing a slight incentive to form larger cliques, these alters are given no prioritization in the search nor do they gain any additional homophily or distance flexibility. Future implementations will focus on strategies to align the clique distributions shown in Figure 9.

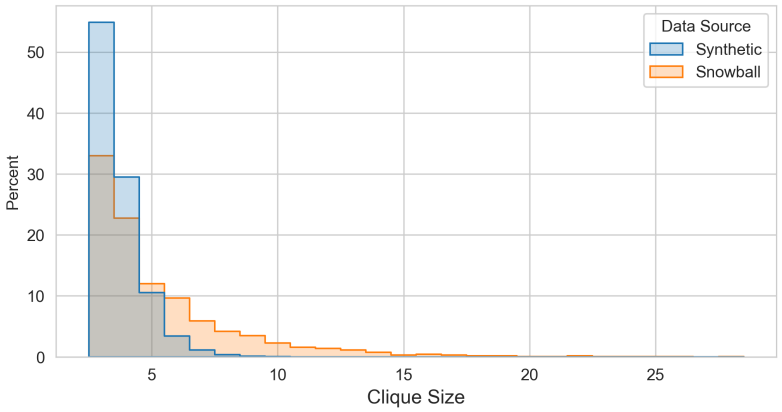

Figure 9: Clique size distribution.

Preferential Attachment Mechanism

Thanks to seminal work by Barabási & Albert (1999), many methodologies for social network formation rely on the concept of preferential attachment. Preferential attachment refers to the tendency of new nodes to connect to high-degree nodes during network formation or evolution. This gives many network degree distributions long right tails. While the snowball data, which limits respondents to 40 names, does not reflect this type of 'scale-free' distribution, the formation model still exhibits degree correlation. Figure 10 demonstrates a positive correlation between an agent's degree and the average degree of its neighbors. This likely emerges from the algorithm's iterative nature. Higher-degree agents generally remain in the algorithm for more iterations; the further into the simulation they get, the more often they encounter other agents with a similarly large degree goal.

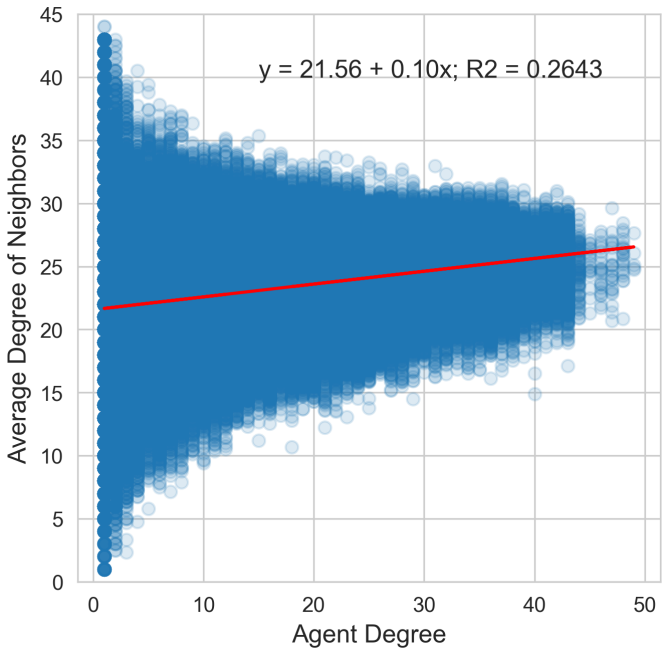

Figure 10: Agent degree versus average neighbor degree.

4 Conclusions

This paper presents a scalable method for generating a synthetic social network with relevant socio-demographic and geospatial characteristics based on a small, egocentric data sample. The generated network demonstrates homophily matching, has clique structure and shows preferential

attachment. A point of improvement would be better clique formation, as the current approach does not result in a clique distribution similar to the input data. Obeying egocentric homophilies while forming very large cliques is statistically unlikely in the current, agent-based framework; perhaps this approach can be blended with the clique-centric formation strategies of Dubernet (2017). Literature in network formation also suggests options related to clustering and community detection (Girvan & Newman, 2002). Additionally, the generated network is static, representing one point in time. Future research could focus on dynamic updating of the social network over time. Lastly, further work on travel behavior analysis should be conducted to assess how social networks influence joint travel decisions, as 45% of trips in Germany were performed with at least one companion (Infas et al., 2018). Further research in synthetic social network generation is a necessary step toward modeling the influence of social networks on travel behavior.

Acknowledgements

The research was funded by the German Research Foundation (DFG) under the research project TENGOS: "Transport and Epidemic Networks: Graphs, Optimization and Simulation" (Project number 458548755).

References

Arentze, T., Kowald, M., & Axhausen, K. (2012). A method to model population-wide social networks for large scale activity-travel micro-simulations. Retrieved from `https://doi.org/10.3929/ethz-b-000038374` doi: 10.3929/ethz-b-000038374

Asikainen, A., Iñiguez, G., Ureña-Carrión, J., Kaski, K., & Kivelä, M. (2020). Cumulative effects of triadic closure and homophily in social networks. *Science Advances*, *6*(19). Retrieved from `https://www.science.org/doi/abs/10.1126/sciadv.aax7310` doi: 10.1126/sciadv.aax7310

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509-512. Retrieved from `https://www.science.org/doi/abs/10.1126/science.286.5439.509` doi: 10.1126/science.286.5439.509

Carrasco, J. A., Hogan, B., Wellman, B., & Miller, E. J. (2008). Collecting social network data to study social activity-travel behavior: An egocentric approach. *Environment and Planning B: Planning and Design*, *35*, 961-980. doi: 10.1068/b3317t

Dubernet, T. (2017). Explicitly correlating agent's daily plans in a multiagent transport simulation: Towards the consideration of social relationships. Retrieved from `https://doi.org/10.3929/ethz-b-000165685` doi: 10.3929/ethz-b-000165685

Frei, A., & Axhausen, K. W. (2007). Size and structure of social network geographies. Retrieved from `https://doi.org/10.3929/ethz-a-005562753` doi: 10.3929/ethz-a-005562753

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821-7826. Retrieved from `https://www.pnas.org/doi/abs/10.1073/pnas.122653799` doi: 10.1073/pnas.122653799

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th python in science conference* (p. 11 - 15). Pasadena, CA USA.

Illenberger, J., Nagel, K., & Flötteröd, G. (2013). The role of spatial interaction in social networks. *Networks and Spatial Economics*, *13*(3), 255-282. Retrieved from `https://doi.org/10.1007/s11067-012-9180-4` doi: 10.1007/s11067-012-9180-4

Infas, DLR, IBT, & infas 360. (2018). *Mobilität in deutschland.* Retrieved 14.06.2019, from `http://www.mobilitaet-in-deutschland.de/publikationen2017.html`

Kim, J., Rasouli, S., & Timmermans, H. J. (2018, 7). Social networks, social influence and activity-travel behaviour: a review of models and empirical evidence. *Transport Reviews*, *38*, 499-523. doi: 10.1080/01441647.2017.1351500

Kowald, M., & Axhausen, K. (2012). Focusing on connected personal leisure networks: Selected results from a snowball sample. *Environment and Planning A: Economy and Space*, *44*(5), 1085-1100. Retrieved from `https://doi.org/10.1068/a43458` doi: 10.1068/a43458

Moreno, A., & Moeckel, R. (2018). Population Synthesis Handling Three Geographical Resolutions. *ISPRS International Journal of Geo-Information*, *7*(5), 174. doi: 10.3390/ijgi7050174