Considering baseline homophily when generating spatial social networks

Sascha Holzhauer[1], **Friedrich Krebs**[1] **and Andreas Ernst**[1]

Abstract. Social networks have become an important part of agent-based models, and their structure may have remarkable impact on simulation results.

We propose a simple but powerful approach for spatial agent based models which explicitly takes into account restrictions and opportunities imposed by effects of baseline homophily, i.e. the tendency to build up relationships with others that are similar. The resulting network thus reflects social settings and furthermore allows the modeller to influence network properties by adjusting agent type specific parameters. Especially the maximum extension of the search radius and the value by which the radius is extended allows for control of clustering and agent type distribution of personal networks.

1 MOTIVATION

The generation of social networks is an important issue in agent-based modelling. The network structure might have considerable impact on certain processes like opinion formation [5], information exchange for problem solving [10], or advice [22]. Furthermore, [4] investigates the impact of network structure in a model of racial segregation and comes to the conclusion that the structure of the social network, and especially its relation to physical space, has significant effects on the results of social simulation.

Usually, simulations generate social networks according either to the small world algorithm proposed by [21] ([8; 9]) or to preferential attachment [1]. These methods focus on producing networks whose global, i.e. network level properties like average path length, clustering coefficient, and degree distribution are as similar as possible to empirically found values. However, these methods neglect local circumstances as well as actor properties and preferences and/or require global network knowledge during the generation. Whereas such aspects may be insignificant with respect to rather theoretical applications they might play a key role in many social science simulations, for instance in modelling for policy consulting.

Social networks are mostly characterised by what is often called the homophily principle. That is, people tend to build up relationships with others that are similar in some or many personal and socio-demographic attributes like age, gender, ethnic origin, educational background, or income. Thus, homophily narrows the people's social world in a fundamental way and influences their access to information, the way they are forming attitudes and the persons they meet. [11] distinguishes between baseline and inbreeding homophily. The former describes the phenomenon that people often live in surroundings with similar others. Consequently, the chance to spend time with that group and build up acquaintances is higher because of the composition of potential others. As a result, more trust occurs in such groups of similar people and network flows of information may increase. On the other hand, barriers between groups may exist which hinder information to spread [19]. The latter term describes the explicit tendency for persons to choose friends that have similar views, related occupations and like the same hobbies above the opportunity set.

[2] present an elaborated model based on social distance attachment that takes inbreeding homophily into account. The probability to link is derived from the sum of distances between individuals regarding each value of a vector representing the individuals' social coordinates. Resulting networks are compared with empirical data of the PGP (pretty good privacy) web of trust, and convincing similarity is obtained with respect to assortativity (i.e., the tendency that highly connected persons tend to have links to others with a high degree and vice versa) and a hierarchical community structure. However, the authors do not consider asymmetrical relationships. One way to accomplish directed networks is to define an individual's position in the social space for both in-going and out-going links.

[6] proposes a network generation method based on social circles [16]. Similar to [2] agents are located on a kind of social map according to certain, e.g. socio-demographic, properties. Whereas [2] proposes a city-block based distance measure (L_1) [6] applies an Euclidean based measure (L_2). Agents whose so-called reaches of a specific radius around their position on the map match each other's get connected. Again, this approach is not suitable for asymmetrical relationships. Furthermore, whereas it is possible to reflect agent specific ego network sizes by different reaches the two-dimensional map does not allow for placing agents according to more than two properties.

For his agent-based simulation [20] accounts for inbreeding homophily tendencies and connects agents according to their network preferences, i.e. the number of desired relationships and the liking either for similar or sometimes even dissimilar persons. The author further discriminates between normative, i.e. influencing, and informational ties. Finally, deviations are defined with respect to the number of relationships, the amount of correct relation types, and the number of desired similar and dissimilar ties. Agents then shall be connected in ways that minimise these deviations.

We propose a network generation process that takes into account baseline as well as inbreeding homophily. Since we build up a spatially explicit social simulation we are mainly interested in the spatial restrictions and opportunities actors face when they make up relationships. An actor may only connect to those others who are available within the boundaries he is agitating. For instance, the choice of network partners may vary

[1] Center for Environmental Systems Research, University of Kassel, Germany. Email: {holzhauer, krebs, ernst}@usf.uni-kassel.de

whether someone lives in a dense urban environment with manifold others to choose from, or in a sparsely populated rural area.

2 OUR APPROACH

An important and comprehensive source of heterogeneity of people is their grouping according to sociological lifestyles. Because of societal liberalisation social norms based on social classes decay and individuals experience more autonomy. Lifestyles seek to capture perceivable patterns of behaviour, symbolic integration and underlying orientations as expressions of that autonomy. Lifestyles are thus meant to be a more relevant grouping of individuals and households [18]. We apply the Sinus-Milieus® [17] that are commonly used in commercial market research, but also in environmental research [14]. Sinus-Milieus® group individuals or households along the classical dimension of social status given by income and education, and supplement this grouping by a second dimension that reflects social value orientations like tradition, modernisation and re-orientation.

The empirical base for the results presented in this paper is a dataset of spatially referenced socio-demographic data of the target region of Northern Hesse located in the centre of Germany. Data originate from a 2007 survey by Microm® [12]. The geographical reference units are cells that comprise one to several hundred households depending on population density. For each of the cells we extract the number of households belonging to each of four different lifestyles: Leading lifestyles are characterised by the pursuit of prestige as well as wealth and occupation of leadership positions. Traditional lifestyles are often adopted by worker families that desire security and order. The mainstream strives for professional and societal establishment and harmonic circumstance, whereas a hedonistic lifestyle is characterised by the search for pleasure, sometimes with little resources, and often the denial of conventions.

In order to apply our network generation approach considering baseline and inbreeding homophily we first initialise an agent population such that the distribution of lifestyles among agents and the agents' location reflect the empirically observed spatial distribution of lifestyles. To do so we first determine the number of required representatives for each lifestyle in every data cell. Then, we initialise each agent as a representative for ten households of a specific lifestyle and place it normally randomly close to the respective cell in a GIS (see figure 1). The resulting population setup is empirically founded and provides spatial relationships between agents as well as lifestyle heterogeneity.

Since we are interested in processes of social influence we model relationships between agents as asymmetrical ties that are represented by directed links in a network. These links have their origin in the influencer and lead to the agent that is being influenced. Therefore, the in-degree of an agent's personal network (also referred to as ego network) specifies the number of network partners that influence that agent. Table 1 presents the lifestyle specific network preferences.

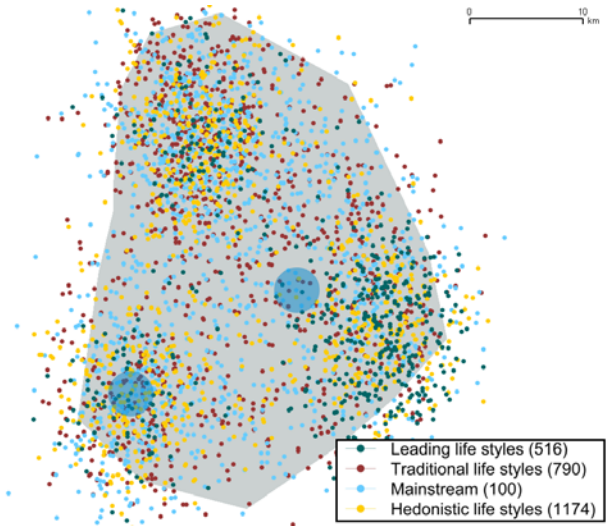

Figure 1: Points represent agent positions within the model region whereas colors specify the agent's lifestyle. Numbers in brackets are the amount of agents of that lifestyle. The total number of agents within the model region is 3480. Cumulations indicate three smaller cities. Blue shaded circles show a search radius of 2000m around an agent in rural area and an agent within a city.

	Leading	Traditional	Mainstream	Hedonistic
In-degree	15	5	5	10
p_rewire	0.2	0.05	0.1	0.2
p_links to				
Leading	0.8	0.0	0.0	0.2
Traditional	0.6	0.3	0.1	0.0
Main-stream	0.6	0.1	0.3	0.0
Hedonistic	0.5	0.0	0.0	0.5

Table 1: Expert rating of lifestyle network preferences. Whereas members of leading and hedonistic lifestyles have far reaching networks and thus are assigned a high rewiring probability, people of traditional lifestyles do not. Data is based on [15].

The network generation is divided into two parts, the establishment of local links and the rewiring process. Each single part is processed iteratively for all agents. As depicted in figure Figure 2 the first part starts with collecting and shuffling all agents within the current search radius which is initially given by START_SEARCH_RADIUS. For every potential partner that is not yet connected with the focal agent it is decided according to the lifestyle specific probability (see p_links in table 1) if it should be linked to the focal agent.

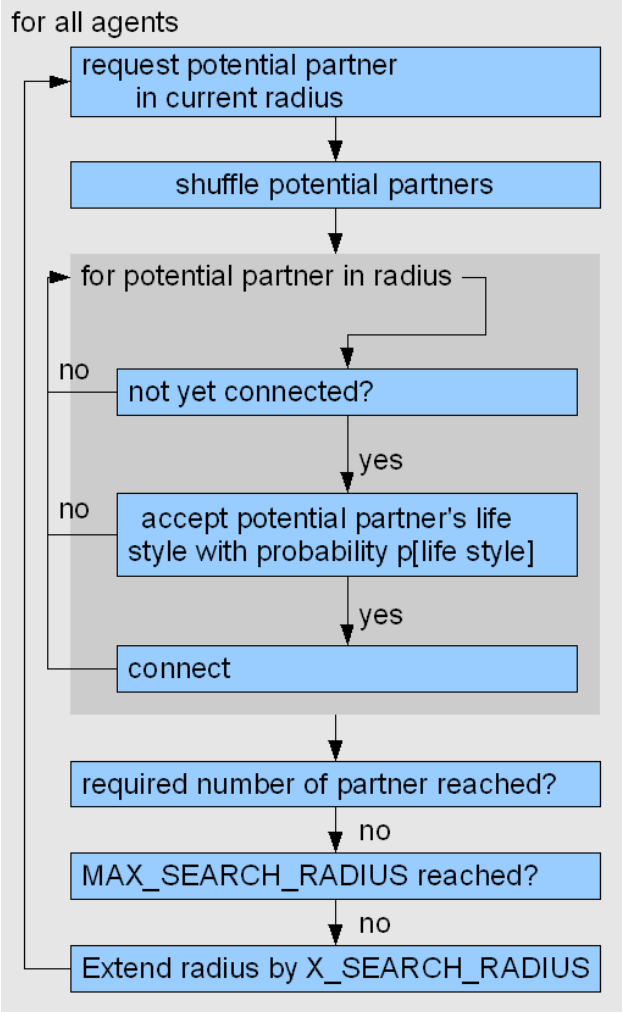

Figure 2: Course of local network generation (rewiring not included)

If the number of required network partners is not reached but all collected agents are treated, more agents are collected from around the focal agent within a current radius that is extended by X_SEARCH_RADIUS. This loop is repeated until either the number of required network partners is satisfied or maximum radius (MAX_SEARCH_RADIUS) is reached.

The approach to select surrounding agents as they come considers the local lifestyle composition and reflects baseline homophily. However, this way the algorithm accounts not only for groups of similar agents that stick together but also for opposite situations when one cannot establish connections to those people one would like to. Applying lifestyle specific preference probabilities when accepting or rejecting a potential network partner reflects inbreeding homophily finally.

After each agent is connected locally the global rewiring process takes place during the second part. For each agent and every existing local link, with probability P_REWIRE (see p_rewire in table 1) the link is rewired to a randomly chosen agent from the entire model region. The random target agent selection is repeated until the found agent is accepted according to the lifestyle specific preferences probabilities (p_links).

The emerging distant links result in the small world effect with high clustering and low average path lengths. On purpose the new partner's lifestyle needs not to be the same as that of the originally linked: The composition of network partners within direct surroundings is characterised by the local lifestyle distribution (baseline homophily) and therefore does not entirely reflect the focal agent's network partner preferences (p_links). Determining the lifestyle during rewiring anew may correct this lifestyle composition of network partners towards imbreeeding homophily and thus is desired.

3 RESULTS

We implemented our spatial agent-based model in Repast Simphony [7]. Data is exported to a database and processed by R [13; 3]. Results are averaged over five independent model runs with different random seed.

We compare the results of our proposed algorithm that takes baseline homophily into account with an ideal network builder and a small world generator [21]. The ideal network builder tracks the lifestyle of network partners and allows a link between the focal agent and a potential alter only if the focal agents has not yet built enough connections to other representatives of the alter’s lifestyle.

To evaluate the appropriateness of certain algorithm variations and parameter settings we introduce some quality measures. The deviation from preferred lifestyle distribution of partners (preference deviation) compares the desired personal network's lifestyle composition with the actual one. The measure sums up the deviation for each of the four lifestyles. The deviation from preferred in-degree to the actual number of influencing others is referenced to as in-degree deviation. Furthermore, we consider the average path length (average network distance of all node pairs in the network) and the global clustering coefficient, also known as transitivity index, which in our case is the number of all existing triples divided by the total number of triangles, i.e. potential triples.

It is important to note that the measures highly depend on the distribution of agents across the model region, especially with respect to lifestyles. Our model region as depicted in figure 1 is a rather rural area with three small cities. For agents in the centre of the area it will be quite hard to satisfy their links with respect to inbreeding homophily. This is especially true for people of leading lifestyles that occur very sparsely in the centre but like to connect predominantly to other people of a leading lifestyle.

There are some parameters to adjust the network's characteristics. Whereas the MAX_SEARCH_RADIUS defines the geographical area within which agents may search for partners, X_SEARCH_RADIUS denotes the value by which the search radius is extended in case the current radius is not far enough to fulfil the number of partners the agent desires. Furthermore, the rewiring probability influences the amount of rather distant links.

Figure 3 shows the network in-degree deviation as a function of MAX_SEARCH_RADIUS. The smaller the radius, the less space is given to fulfil the agents’ preferences regarding the lifestyle distribution of their social network. The algorithm considering baseline homophily yields lower deviations for larger radii since it allows connections to alteri that do not match the preferred lifestyle distribution. Of course, regarding network preference deviation the ideal network builder performs better since that is its purpose. As figure 4 clearly indicates, with increasing MAX_SEARCH_RADIUS the deviations can be reduced. Leading lifestyles improve only slightly since the overall number within the model region is limited. In terms of modelling realistic social networks a specific deviation is desired for certain lifestyles since it reflects social settings.

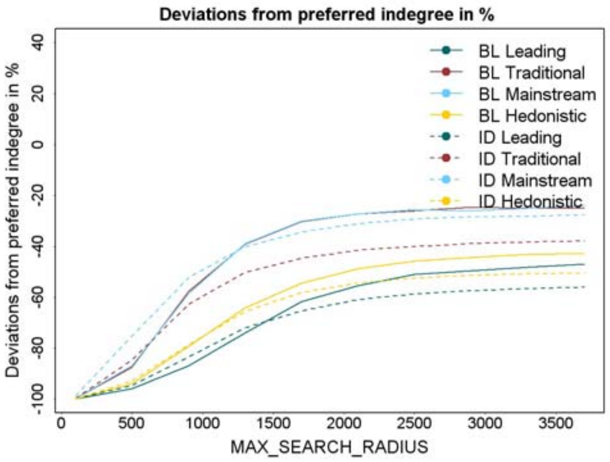

Figure 3: Percental Network in-degree deviation with raising MAX_SEARCH_RADIUS. Negative values indicate that actual degree is smaller than preferred. For smaller radii, the algorithm considering baseline homophily (BL – dashed lines) yields higher deviations from the preferred in-degree (number of influencers) than the ideal network builder (ID – solid lines).

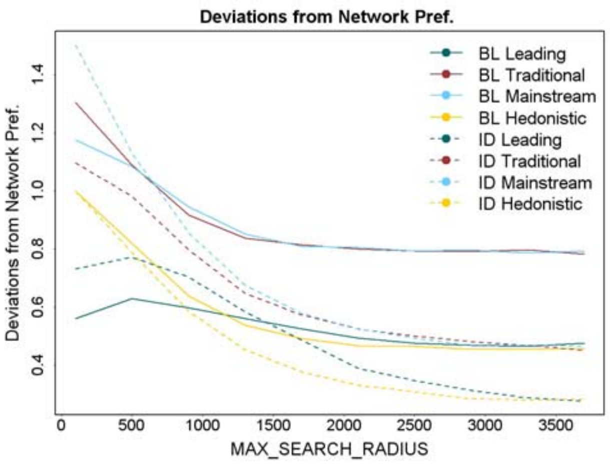

Figure 4: Deviations from milieu-specific network partner preferences with raising MAX_SEARCH_ RADIUS. Compared to the ideal network builder (ID-solid lines) the baseline algorithm (BL – dashed lines) results in higher deviations (apart from hedonists). With increasing MAX_SEARCH_RADIUS deviations become smaller.

As figure 5 shows, the average distance to a neighbour is considerably lower in networks from the proposed builder. Of course, this is due to the local search for neighbours the small world generator does not take into account. As the rewiring probability raises also the average distance increases.

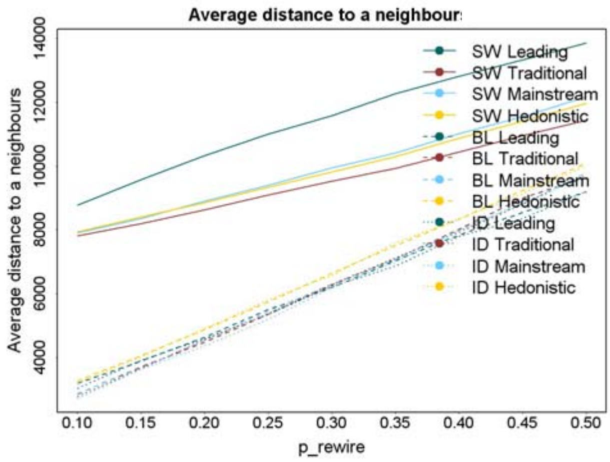

Figure 5: Average distance to a neighbour in meters. Since the small world generator does not explicitly consider spatial proximity, the distance is larger.

The rewiring probability relaxes the MAX_SEARCH_RADIUS in the way that it allows the agents to choose the more agents deliberatively within the entire simulation area the higher the probability is. Furthermore, it is in particular responsible for the small world properties and thus affects the average path length and the clustering coefficient. The global clustering coefficient gives an important hint towards the empirical foundation of the proposed network generation algorithm. The higher the amount of local links that are rewired globally the lower is the clustering coefficient (see figure 6) length and lower is the average path (see figure 7) [21].

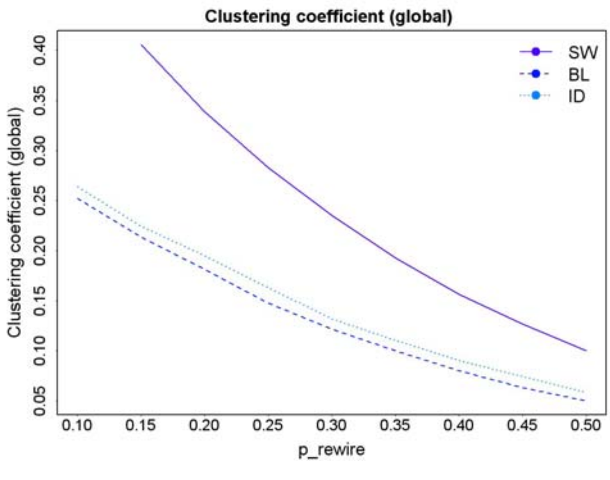

Figure 6: The global clustering coefficient drops strongly when more and more local links are globally rewired. The small world generator yields a much higher clustering.

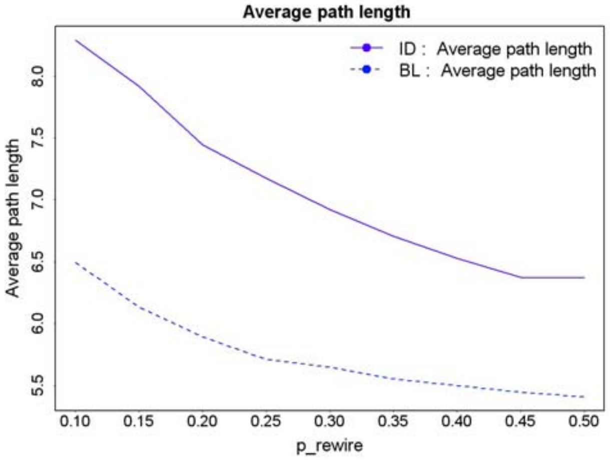

Figure 7: The Average path length decreases along with the establishment of more distant relationships.

As figure 8 indicates, variations in the rewiring probability have also a minor impact on the network preference deviations. Whereas for the proposed algorithm deviations decrease because rewiring guarantees a partner of desired lifestyle, network produced by the ideal network builder do not benefit from rewiring. That is because the target agent is not guaranteed to be of the desired lifestyle.

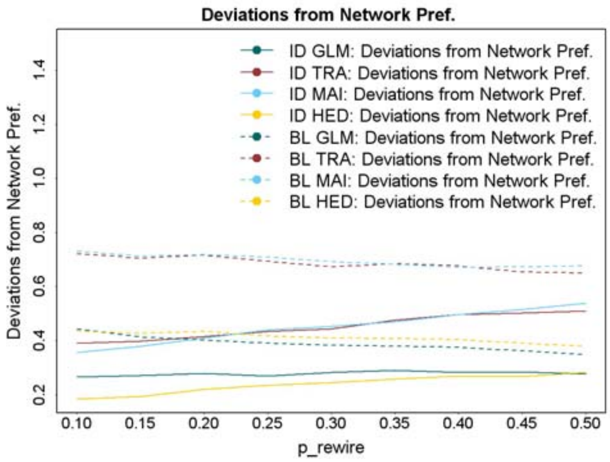

Figure 8: For the baseline homophily considering approach, deviations from preferred lifestyle distribution of network partners decrease with increasing rewiring probability since rewiring supports partners of desired lifestyle.

Figure 9 shows the effect of altering the X_SEARCH_RADIUS, that is the radius by which the search radius is extended in case the number of required partners can not be fulfilled, has on the clustering coefficient. If the search radius is raised slowly, agents are forced to build up connections with nearby agents which supports local clustering. However, since a smaller search radius reduces the opportunity set, the network preferences deviation is lower for higher values for X_SEARCH_RADIUS.

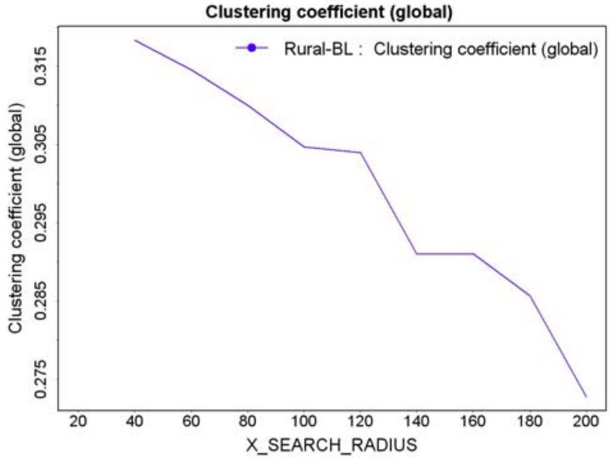

Figure 9: Raising X_SEARCH_RADIUS when the initial search radius is rather small (20m). The clustering coefficient is higher for small values of X_SEARCH_RADIUS when agents are forced to build up rather local connections.

Finally, we investigate the impact of the baseline homophily considering approach on the out-degree distribution, i.e. the number of network partner a focal agent may influence. Compared to the ideal network builder agents are assigned more outgoing relationships. The reason is that the baseline homophily concept is less strict in the selection of alteri. Leading lifestyles (dark green line) are especially central in the network (figure 10).

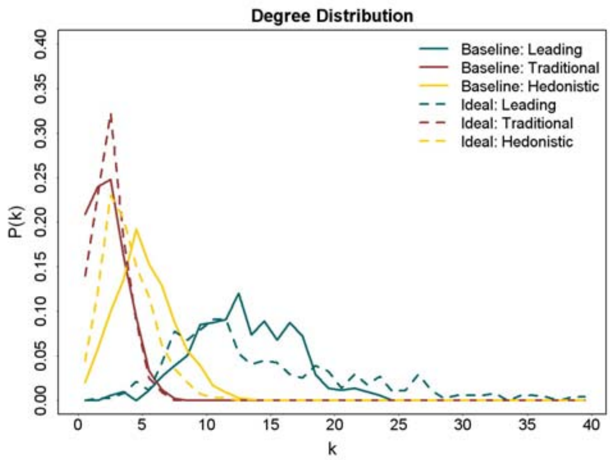

Figure 10: Distribution of out-degree for MAX_SEARCH_RADIUS of 2500m and X_SEARCH_RADIUS of 100m. Since the baseline algorithm is more flexible in assigning partners degree distributions are shifted to the right.

In comparing the baseline homophily considering network generator with a small world generator we find that the latter yields somewhat smoother network properties (e.g., see clustering coefficient in figure 6). However, taking the principles of baseline homophily into account might question the realism of that widespread network generator's foundation. As figure 11 shows, the proposed network generator results in moderate assortative mixing, due to local restriction in partner selection.

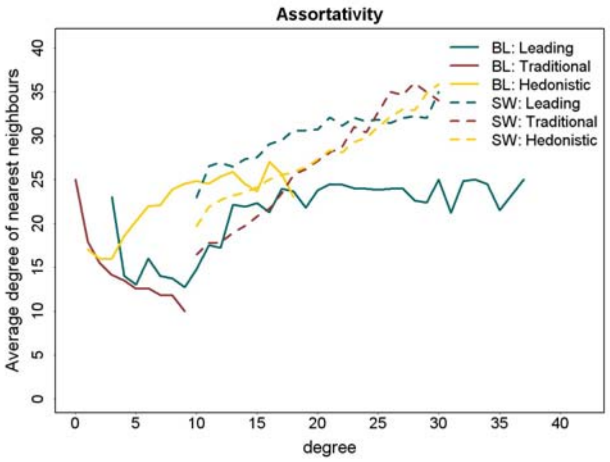

Figure 11: The average degree of nearest neighbours as a function of degree shows moderate assortative mixing (MAX_SEARCH_RADIUS: 3000m, rewiring of 0.1, X_SEARCH_RADIUS: 500m).

4 DISCUSSION

We proposed a simple but powerful approach to generate social networks for spatial agent based models. It seeks to reflect realistic, natural settings of the model region and also shows desired, empirically grounded network properties like short average path length, considerable clustering, and moderate assortativity. Therefore, we describe an alternative to the widespread small world algorithm which lacks realistic groundings with respect to local interactions.

The resulting network may be adjusted by setting the MAX_SEARCH_RADIUS (to set the moving radius of actors which might differ considerably from area to area and from life style to life style), the X_SEARCH_RADIUS (the radius by which the search radius is extended as long as more agents are required to choose from and MAX_SEARCH_RADIUS is not reached), and the P_REWIRE (to account for network parts that outreach the local region). Furthermore, the lifestyle preferences of each agent type may be adjusted. The MAX_SEARCH_RADIUS provides an adequate regulator to adjust milieu-specific radii of action and thus reduce the network preference deviation while preserving clustering. X_SEARCH_RADIUS helps to control the clustering coefficient, while p_rewire has an impact on the average path length.

Probably the greatest challenge in modelling social networks is gaining adequate empirical data about the relations modelled actors have. An advantage of our approach is that every parameter could be more or less empirically measured. For instance, the MAX_SEARCH_RADIUS is determined by the area a person normally agitates within. The network size and preferences regarding life styles could be gained by analysing personal networks of an adequate amount of representatives of each life style. However, since such explorations are quite demanding and expensive one most often has to guess values from experience or consult experts in the field.

In the future we seek to further explore the parameter space of the network generation in order to predict the properties of resulting networks more thoroughly. Emphasis is placed on the interplay between the mentioned parameters. For instance, both the rewiring probability and X_SEARCH:_RADIUS have an impact on the global clustering coefficient. Besides it is worth to explore heterogeneous, lifestyle specific parameters.

A possible extension is to allow agents to start their search within a specific annulus around their home coordinates and then broaden it simultaneously to the inner and the outer area. This would account for people that refuse to make connections within their direct neighbourhood. Furthermore, extensions in the direction of incorporating geographical and social distance as proposed by [2] is expected to be fruitful.

REFERENCES

[1] A.-L. Barabasi, and R. Albert: Emergence of Scaling in Random Networks, *Science* 286 (1999) 509–512

[2] M. Boguna, R. Pastor-Satorras, A. Diaz-Guilera, and A. Arenas: Models of social networks based on social distance attachment, *PHYSICAL REVIEW E* 70 (2004)

[3] G. Csardi, and T. Nepusz: The igraph software package for complex network research, *InterJournal* Complex Systems (2006) 1695

[4] B. Edmonds: How are physical and social spaces related? Cognitive agents as the necessary glue. In: *Agent-Based Computational Modelling*, F.C. Billari, T. Fent, A. Prskawetz, and J. Scheffran, Eds.: SPRINGER (2006)

[5] A. Grabowski: Opinion formation in a social network: The role of human activity, *Physica A - Statistical Machanics and its Applications* 388 (2009) 961–966

[6] L. Hamill, and N. Gilbert: Social Circles: A Simple Structure for Agent-Based Social Network Models, *Journal of Artificial Societies and Social Simulation* 12 (2009) 3

[7] T.R. Howe, N.T. Collier, M.J. North, M.T. Parker, and J.R. Vos: Containing Agents: Contexts, Projections, and Agents. In: *Proceedings of the Agent 2006 Conference on Social Agents: Results and Prospects*, Argonne, Illinois, USA (2006)

[8] M.A. Janssen, and W. Jager: Stimulating diffusion of green products - Co-evolution between firms and consumers, *Journal of Evolutionary Economics* 12 (2002) 283–306

[9] L.-L Jiang, D.-Y Hua, J.-F Zhu, B.-H Wang, and T. Zhou: Opinion dynamics on directed small-world networks, *European Physical Journal B* 65 (2008) 251–255

[10] D. Lazer, and A. Friedman: The network structure of exploration and exploitation, *Administrative Science Quarterly* 52 (2007) 667–694

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

[12] Micromarketing Systeme und Consult GmbH, *Microm Consumer Marketing* (2011), Available at http://www.microm-online.de.

[13] R Development Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria (2010), Available at http://www.R-project.org/.

[14] N. Schwarz, and A. Ernst: Using empirical data to build an agent-based model of innovation diffusion. In: *Proceedings of the workshop on agent-based models of market dynamics and consumer behaviour,* Surrey, GB (2006)

[15] N. Schwarz, *Umweltinnovationen und Lebensstile: Eine raumbezogene, empirisch fundierte Multi-Agenten-Simulation.* 1st ed.: Metropolis Verlag (2007)

[16] G. Simmel: The number of members as determining the sociological form of the group, *American Journal Of Sociology* 8 (1902) 1–46

[17] Sinus Institut, *Sinus-Milieus* (2011), Available at http://www.sinus-institut.de/loesungen/sinus-milieus.html.

[18] A. Spellerberg, *Soziale Differenzierung durch Lebensstile*. Berlin: Edition Sigma (1996)

[19] Thomas Valente, *Social Networks and Health: Models, Methods, and Applications.* 1st ed.: Oxford University Press Inc (2010)

[20] P. van Eck: Social Network Structures in Agent Based Modeling: Finding an Optimal Structure Based on Survey Data (or Finding the Network That Does Not Exist). In: *Proceedings of the 3rd World Congress on Social Simulation WCSS2010*, A. Ernst, and S. Kuhn, Eds., Kassel, Germany (2010)

[21] D.J. Watts, and S.H. Strogatz: Collective dynamics of ‘small-world‘ networks, *Nature* 393 (1998) 440–442

[22] S.-S. Wong: Judgment about knowledge importance: the roles of social referents and network structure, *Human Relations* 61 (2008) 1565–1591