Complex System project: Homophily & Heterophily on Human Interaction

Udomlerd SRISUCHINWONG & Nattapon PREEDASAK

Abstract

We study the homophily & heterophily in human attributes (e.g. age, seniority) by modeling the clustering networks. This work aims to investigate how to quantify and interpret homophily & heterophily based on the data from the International Conference on Computational Social Science 2017 (ICCSS17)[5]. In our model, we apply the stochastic block model (SBM) network with homophily parameters to compute the homophily & heterophily for each attribute by contact matrix. We find that the large number of nodes in the same node group is a main role to dominate the homophily, while the small number of nodes in the node group is more likely to disperse to other groups (i.e. heterophily) for all attributes.

1 Introduction : Homophily & Heterophily

Homophily: The word *homophily* was coined to bring together the observations of early network researchers and relate them to classic anthropological studies of homogamy (homophily in marriage formation). McPherson mentioned that

> *Homophily* is the principle that a contact between similar people occurs at a higher rate than among dissimilar people. The pervasive fact of homophily means that cultural, behavioral, genetic, or material information that flows through networks will tend to be localized. Homophily implies that distance in terms of social characteristics translates into network distance, the number of relationships through which a piece of information must travel to connect two individuals. It also implies that any social entity that depends to a substantial degree on networks for its transmission will tend to be localized in social space and will obey certain fundamental dynamics as it interacts with other social entities in an ecology of social forms[10].

According to Lazarsfeld and Merton (1954)[7], we can classify *homophily* into two distinct types: *status homophily* (based on informal, formal, or ascribed status.) and *value homophily* (based on values, attitudes, and beliefs). However, we focus only the *status homophily*. It includes the major sociodemographic dimensions that stratify society-ascribed characteristics like race, ethnicity, sex, or age, and acquired characteristics like religion, education, occupation, or behavior patterns. Here are several examples of homophily: *i*) *Race and ethnicity*: Strong homophily on race and ethnicity may be seen in a wide range of connections, from the most personal bonds of marriage and confiding to the narrow networks of discussion about a specific subject. *ii*) *Sex*: Work establishments, for example, are separated by gender[2]. *iii*) *Age*: Age homophilous relationships are more intimate, last longer (typically reflecting the persistence of childhood bonds), involves a greater number of interactions, and are more personal[4]. *iv*) *Education and social class*: Yamaguchi (1990)[11] reported that homophily in education extended to inbreeding bias among friends' social statuses, with one educational level influencing other educational level choices.

Heterophily: On one hand the notion of *homophily* relates to social features or attributes that shared by groups of people involve to deeper relation behaviors among their members; the *heterophily*'s concept, on the other hand, refers to the presence of common relation behaviors between different groups[9].

Stochastic Block Model (SBM): SBM is a statistical model for analyzing latent cluster structures in network data[1]. It is a generalization of the Erdos-Renyi random graph model with a higher intra-cluster probability and a lower inter-cluster probability. The classic SBM relies solely on network connections to infer community structures. When information on nodes is given, the node's information from data can infer its network structure [8], e.g. two types of node given same attribute tend to connect each other by *homophily*.

2 Methods

2.1 Modeling: Homophily Probability

We take a simple, static, clustering network by stochastic block model (SBM) and add homophily as a parameter to our model. The connecting network in our model is determined by the interaction between the degree of nodes and the homophily. In practice, we start to model networks with different types of node groups, α and β. Then we quantify the *homophily* h (conversely, *heterophily* ϵ), as free tunable parameters. This controls how nodes connect with each other based on their attributes. The homophily parameter h is a dimensionless number that varies from 0 to 1, where $h = 0$ denotes the nodes from one group specifically connecting with nodes from the other group (*heterophily*-dominated), while $h = 1$ indicates the nodes within the same groups exclusively connecting to one another (*homophily*-dominated), and $h \sim 0.5$ means *mixing* connection between those groups. Given all connecting links l, the homophily probability of a node i connecting from node j is expressed as

$$\pi_i = \frac{h_{\alpha\beta}(i,j)k_i}{\sum_l h_{\lambda\beta}(\bar{l},j)k_l}, \tag{1}$$

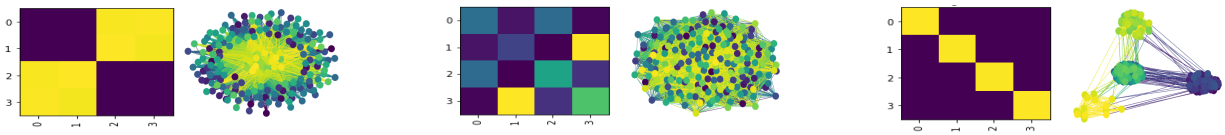
where k_i is degree of the node i, and $h_{\alpha\beta}(i,j) \equiv h$ is homophily parameter of node i connecting node j within or between group α and β. The homophily parameter, in general, determines the probability of connection within and between groups. We classify homophily parameters in different groups: *i)* $h_{\alpha\alpha}$ (probability of connection within members of group α), *ii)* $h_{\beta\beta}$ (probability of connection within members of group β), *iii)* $h_{\alpha\lambda}$ or $h_{\beta\lambda}$ (probabilities of connection between members of group α (or β) and complementary groups λ), and *iv)* $h_{\alpha\beta} = 1 - h_{\alpha\alpha} - \sum_\lambda h_{\alpha\lambda}$; $h_{\beta\alpha} = 1 - h_{\beta\beta} - \sum_\lambda h_{\beta\lambda}$ (probabilities between groups α, β). In our model, we consider homophily as 4-dimensional parameters (4×4 squared matrix). And we assume that homophily is controlled by homophily parameter h, and non-diagonal parameter ϵ. Both are required to be symmetric and complementary i.e. $(h_{\alpha\alpha} = h_{\beta\beta} = h_{\lambda\lambda} = h)$, $(h_{\alpha\lambda} = h_{\beta\lambda} = h_{\lambda\alpha} = h_{\lambda\beta} = \epsilon)$ and $(h_{\alpha\beta} = h_{\beta\alpha} = 1 - h - 2\epsilon) \equiv \Delta$[6]. We will adopt these free parameters as the inputs for 4×4 matrix of probability homophily in the Stochastic Block Model (SBM) network in our model.

Homophily Matrix: To compute homophily probability in SBM, we take homophily matrix $h_{\alpha\beta}$ from eq.(1) as 4×4 symmetric one, and set element parameters h, ϵ, Δ as previously described. It can be expressed in eq.(2)

$$h_{\alpha\beta} = \begin{bmatrix} h_{00} & h_{01} & h_{02} & h_{03} \\ h_{01} & h_{11} & h_{12} & h_{13} \\ h_{02} & h_{12} & h_{22} & h_{23} \\ h_{03} & h_{13} & h_{23} & h_{33} \end{bmatrix} = \begin{bmatrix} h & \Delta & \epsilon & \epsilon \\ \Delta & h & \epsilon & \epsilon \\ \epsilon & \epsilon & h & \Delta \\ \epsilon & \epsilon & \Delta & h \end{bmatrix} \tag{2}$$

Grouping Probabilities: As $h_{\alpha\beta}$ computed, homophily probability of only one node π_i is considered. To compute probability as clustered nodes within group, we assign the grouping probability as the sum of homophily probability of nodes i in group α as P_α. However, in our model we take the grouping probability of homophily between two groups via geometric means[3]. Homophily probability between group α and β is defined as $P_{\alpha\beta}$ in eq.(3). We apply $P_{\alpha\beta}$ as the homophily probability between cluster groups in SBM network for this work.

$$P_\alpha = \left(\sum_i \pi_i \right)_\alpha \qquad \boxed{P_{\alpha\beta} = \frac{(P_\alpha P_\beta)^{1/2}}{(P_\alpha P_\beta)^{1/2} + [(1 - P_\alpha)(1 - P_\beta)]^{1/2}}} \tag{3}$$



(a) Heterophily: $h = 0.001$ **(b)** Mixing : $h = 0.25$ **(c)** Homophily: $h = 0.95$

Figure 1: The visualization of different homophilies with contact matrices and Stochasic Block model (SBM) networks: *heterophily-dominated* ($h \to 0$), *homogeneous-mixing* ($h \sim \epsilon \sim 0.25$), and *homophiliy-dominated* ($h \to 1$) respectively.

2.2 Procedure

In our model we assign a non-directed, clustering network weighted by homophily parameter for each attribute as a Stochastic Block Model (SBM) network. We apply the data attribute from 262 participants in `metadata` file of ICCSS17, which contains *age, sex, country, language, seniority, background, role*, and *previous attendance* attribute for each participant. We obtain partial socio-demographic 202 participants, while 188 with all socio-demographic information[5]. In this section, we describe the basic idea of our procedure as presented in Figure 2 for this project. Further details on the computation in each step can be found in the code file.

1. **Data attribute imputation:** The participant list with non-given data (`NaN`) is replaced with new values. We investigate to apply a method to substitute certain missing data for all socio-demographic participants, by employing a *mode imputation* to replace the missing values for simplicity in each attribute.

2. **Degree extraction:** In parallel, we extract the node degree k_i from temporal data `tij` of ICCSS17 at each time for each node. The degree k_i is later used to compute the homophily probability of each node.

3. **Node group clustering:** After imputed the missing attribute data, we divide each attribute into 4 node groups, namely n_{g0}, n_{g1}, n_{g2} and n_{g3}. Each group represents the number of nodes in each cluster in SBM.

4. **Random partition network:** Given homophily matrix $h_{\alpha\beta}$ & node groups n_g, we build random partition graph to compute homophily parameter in each pair of node in each group from a generated network.

5. **Homophily probabilities:** Combining degree k_i from data and homophily parameters in each pair, homophily probability for each node is computed by eq.(1). To cluster the probabilities of all nodes within/between groups, we compute the average grouping probability by eq.(3). We obtain the mean probabilities within group (diagonal elements of 4×4 probability homophily matrix $P_{\alpha\beta}$), and between group (non-diagonal elements of probability homophily matrix $P_{\alpha\beta}$) as homophily probabilities for SBM.

6. **Stochastic block model network:** Applying the node groups n_g (from step 3), and the homophily probabilities within or between groups $P_{\alpha\beta}$, we establish a stochastic block model network from probability homophily representing the homophily (& heterophily) behaviors in different node in each node cluster.

7. **Contact matrices:** To represent homophily correlations between node clusters at specfific attribute, This can be shown by computing homophily (& heterophily) contact matrices from the SBM network [5].

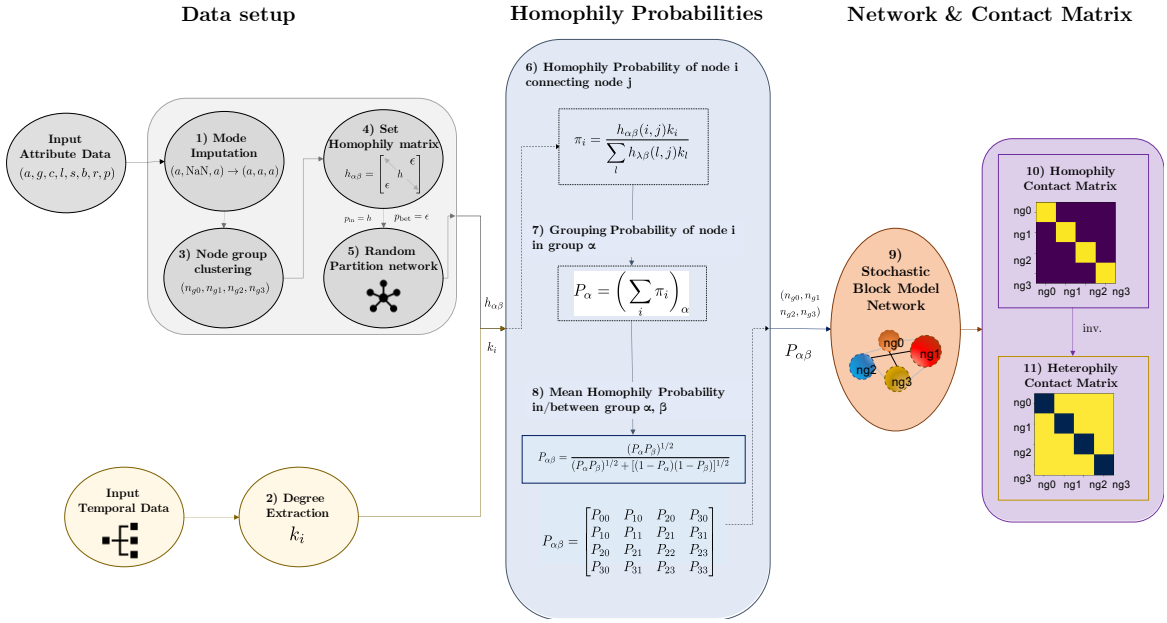


Figure 2: The scheme in our model for the computation of homophily (per attribute) by stochastic block model.

3 Results

For clustering in node groups in each attribute, we classify all attributes into four node groups as shown in Fig. 4 (see the appendix). Our homophily probability within or between group $P_{\alpha\beta}$ is determined by number of nodes in each node group. Certain data in attribute are merged in order to compute 4×4 homophily probability matrix ($P_{\alpha\beta}$). These probabilities result in the likelihood of connection within or between node groups of their attributes, representing the homophily and heterophily of their SBM networks by way of contact matrix (CM).

Homophily is represented by contact matrix. In our model we presume that the sum of contact matrix and inverse contact matrix (ICM) are equivalent to unity (indicating the full probability of connecting network). This implies that heterophily is also represented by inverse contact matrix. Inverse contact matrix is computed by the difference of unity and contact matrix of homophily (i.e. ICM $= 1 -$ CM), implying the heterophily of each attribute in our model. We separate the homophily parameters in three cases: $h = 0.95$, $h = 0.25$, and $h = 0.001$, given heterophily parameter $\epsilon = 0.01, 0.25, 0.333$ respectively. The homophily and heterophily of different attributes by SBM are shown in our work e.g. age homophily in Fig.3, also the seniority, background, country and language homophily (and their heterophily) are presented from Fig. 5 to Fig. 8 in the appendix.

As reference by null models, the random network of Erdos-Renyi model is taken into account with 100 randomisation in order to compare our SBM as the reference models. However we have not fully randomized the networks. We constrain the number of node and number of edge in each network of SBM per homophily parameter and per attribute. The degree of each node in null models is randomized. The deviation of observed values in random distribution is evaluated by z-score [5] as shown in null models for both homophily & heterophily.

Considering SBM's contact matrices in each attribute, our results show that the higher number of nodes in the node group represents the more homophily in the SBM network (thus less heterophily) for all attributes, indicating highly connecting within groups. This corresponds the number distribution of node in different node groups in the Fig. 4, while the lower number of nodes in the node group represents the more heterophily (thus less homophily) in the network of SBM, suggesting highly connecting between groups by inverse contact matrix.

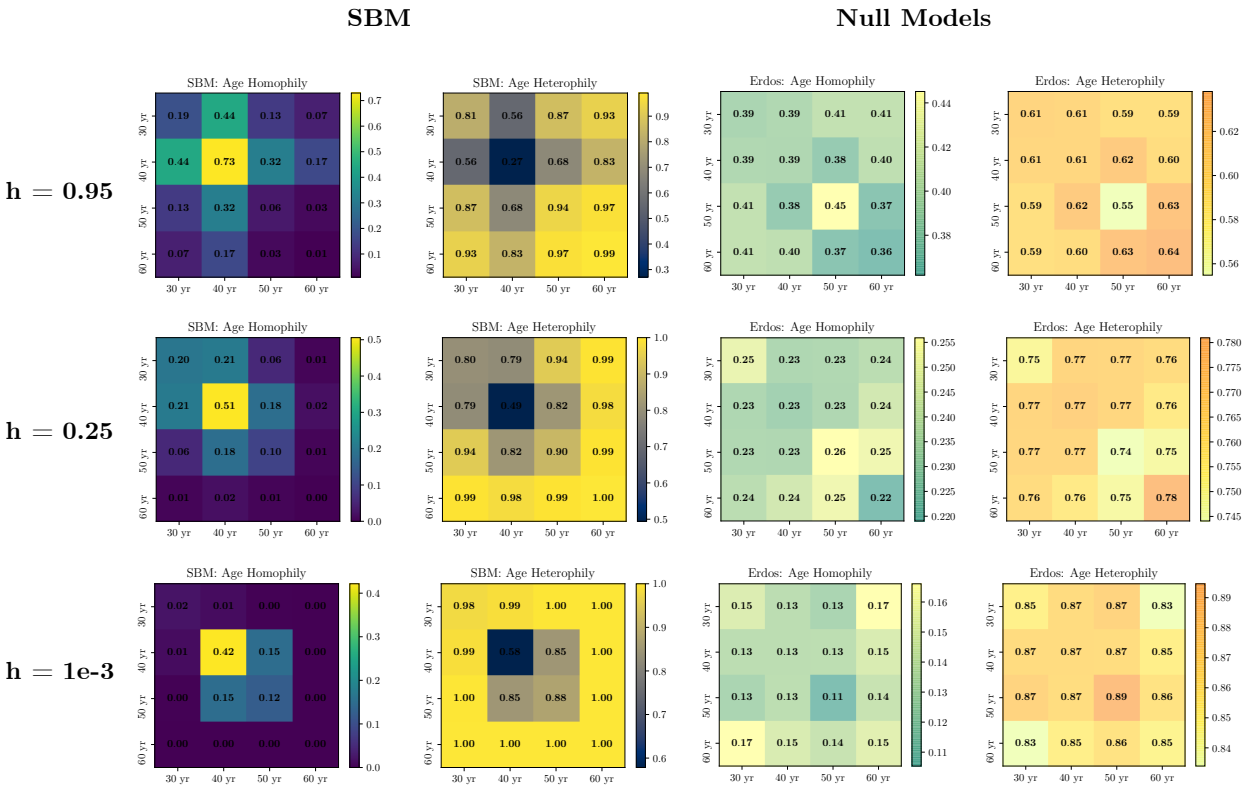


Figure 3: Contact matrices of age homophily via. homophily probability within or between groups. Age homophily for homophily parameter $h = 0.95, 0.25$ and 0.001 (with heterophily parameter $\epsilon = 0.01, 0.25$ and 0.333) respectively is represented from top to bottom contact matrices, where two left columns are stochastic block model and the two right ones are Erdős-Renyi Null model. Age heterophily is calculated by the inverse of age homophily in contact matrices.

4 Discussion

Our work in SBM shows the attributes' homophily indicated by the clustering of same node groups, and their heterophily suggested by the clustering of different groups. Certain attributes are ignored in SBM in this work such as *sex* attribute, *role of conference*, and *previous participation* due to limited number of node groups for 4×4 homophily probability matrix. When the number of nodes in node cluster increases, the probability homophily within groups increases and the probability homophily between groups decreases for most cases. However, as a consequence, the deviation (or variance) comes along with the number of nodes in their clusters. With large homophily parameter h and small heterophily parameter ϵ (i.e. *homophily-dominated*), the more likely nodes are accreting within groups. When compared to null models, the deviations for large homophily parameters are significant for homophily, and are much larger for heterophily due to inversion. Conversely, for high heterophily parameter ϵ with negligible h (i.e. *heterophily-dominated*), a large number of nodes are highly clustered between groups. Their deviations of heterophily by random distribution are far more considerable when compared to higher values h, while homophily deviations are the least with very small number of nodes within groups. From all of these models, these may imply that the more homophily parameter h is given, the variance of distribution is greater in SBM (compared to random network). Further model may require another method to constrain the variance of contact matrices in SBM for high homophily parameters.

The SBM in our model, as we mentioned in the methods, did not consider the direction of connecting nodes in network to compute the probability of link e.g. interaction in Genois et al. (2019) model[5]. Instead, our SBM requires the homophily probability within or between groups $P_{\alpha\beta}$ in symmetric matrix form (as well as homophily matrix $h_{\alpha\beta}$) with the constrain of symmetric index of node groups α, β (e.g. $P_{03} = P_{30}$). In our computation, the geometric mean of the probability of nodes assigned in group α, β is slight difference to that in group β, α due to different degrees and number of edges in their groups. However, in practice, we treat them as equivalence not only for simplicity, but also the requirement of symmetric probability matrix in SBM network. If the directed-network is taken into account in SBM, the computation of homophily probability in eq.(1) might be different. The edges connecting node, and the way of assigning weight in the link by homophily parameter are dependent on time. The homophily probability within or between groups is thus time-dependent. The result of contact matrices depending on time for this model is too sophisticated to represent the homophily and heterophily in each attribute in simple manner, which is beyond our scope for this work.

We have attempted to model the heterophily represented by inverse contact matrix via. matrix operation. By the operation of inverse matrix, it needs to compute the determinant inside contact matrix. However, the determinant of contact matrix occasionally can become zero, if there is a lot of zero elements inside matrix leading to the divergence of inverse contact matrix. Even though the contact matrix's determinant in occasion may not be zero, its inverse matrix still causes the elements with diverged values which are considerably large and greater than unity. Therefore, inverse contact matrix in our model is not considered by matrix operation.

Heterophily parameter ϵ in our work can be only varied from 0 to 0.5 due to complementary parameter Δ that constrains the homophily parameter h within zero to one value. Given that in our SBM network, we neglect the heterophily parameter as a weight on the link for connecting nodes. Only homophily parameter h to assign the weight on our SBM network is taken into account for simplicity. In case heterophily parameter ϵ can be varied through zero to one, heterophily may need other parameter and constrain to quantify itself. For instance, the summation of elements in each row of homophily matrix is more than unity. Nevertheless, further investigation to improve the representation of homophily and heterophily may be required in the future model.

5 Conclusion

In our model, we apply the stochastic block model (SBM) network to represent the homophily & heterophily on the social interaction by considering different attributes: *age, seniority, background, country* and *language*. In particular attribute, the homophily indicates the clustering of the same groups, while the heterophily represents the clustering of different groups. These can be determined by contact matrix and inverse contact matrix, respectively by way of homophily probability within or between groups on the SBM network. The results from our SBM network show that the larger number of nodes in each node group indicates the more homophily behavior (thus less heterophily) in that attribute, representing the highly connecting network within groups, especially in the homophily-dominated system ($h \to 1$). Conversely, the smaller number of nodes in each node group in specific attribute represents the more heterophily behavior (thus less homophily), suggesting the highly connecting network between node groups, particularly in the heterophily-dominated system ($h \to 0$).

References

[1] Emmanuel Abbe. *Community detection and stochastic block models: recent developments*. 2017. arXiv: 1703.10146 [math.PR].

[2] William T Bielby and James N Baron. "Men and women at work: Sex segregation and statistical discrimination". In: *American journal of sociology* 91.4 (1986), pp. 759–799.

[3] Franz Dietrich and Christian List. "Probabilistic opinion pooling". In: (2016).

[4] Claude S Fischer. *To dwell among friends: Personal networks in town and city*. University of chicago Press, 1982.

[5] Mathieu Génois et al. "Building connections: How scientists meet each other during a conference". In: *arXiv preprint arXiv:1901.01182* (2019).

[6] Fariba Karimi et al. "Homophily influences ranking of minorities in social networks". In: *Scientific reports* 8.1 (2018), pp. 1–12.

[7] Paul F. Lazarsfeld and Robert K. Merton. "Friendship as a social process: a substantive and methodological analysis". In: *Freedom and Control in Modern Society* (1954), pp. 18–66.

[8] Yan Leng, Tara Sowrirajan, and Alex Pentland. *Interpretable Stochastic Block Influence Model: measuring social influence among homophilous communities*. 2020. arXiv: 2006.01028 [cs.SI].

[9] Carlos Lozares et al. "Homophily and heterophily in personal networks. From mutual acquaintance to relationship intensity". In: *Quality & Quantity* 48.5 (2014), pp. 2657–2670.

[10] Miller McPherson, Lynn Smith-Lovin, and James M Cook. "Birds of a feather: Homophily in social networks". In: *Annual review of sociology* 27.1 (2001), pp. 415–444.

[11] Kazuo Yamaguchi. "Homophily and social distance in the choice of multiple friends an analysis based on conditionally symmetric log-bilinear association models". In: *Journal of the American Statistical Association* 85.410 (1990), pp. 356–366.

Appendix

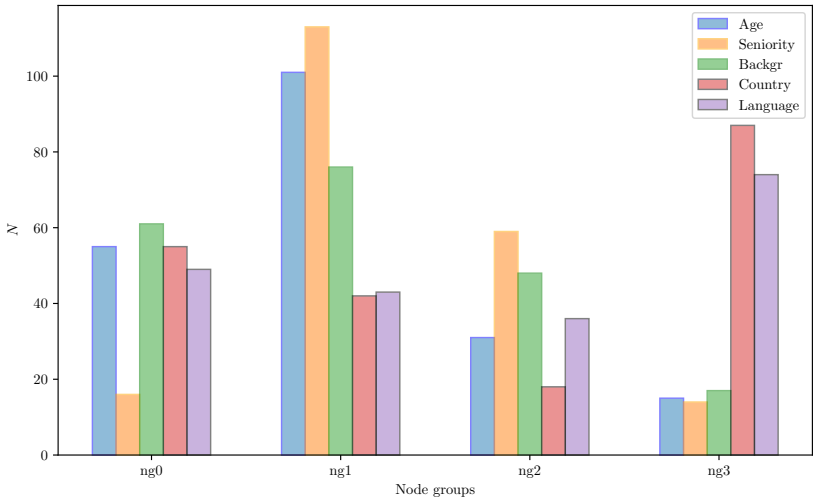


Figure 4: Histogram of node group clusters of human attributes: *age, seniority, academic background, country*, and *language* categorized into four node groups for computing 4×4 probability homophily in stochastic block model.

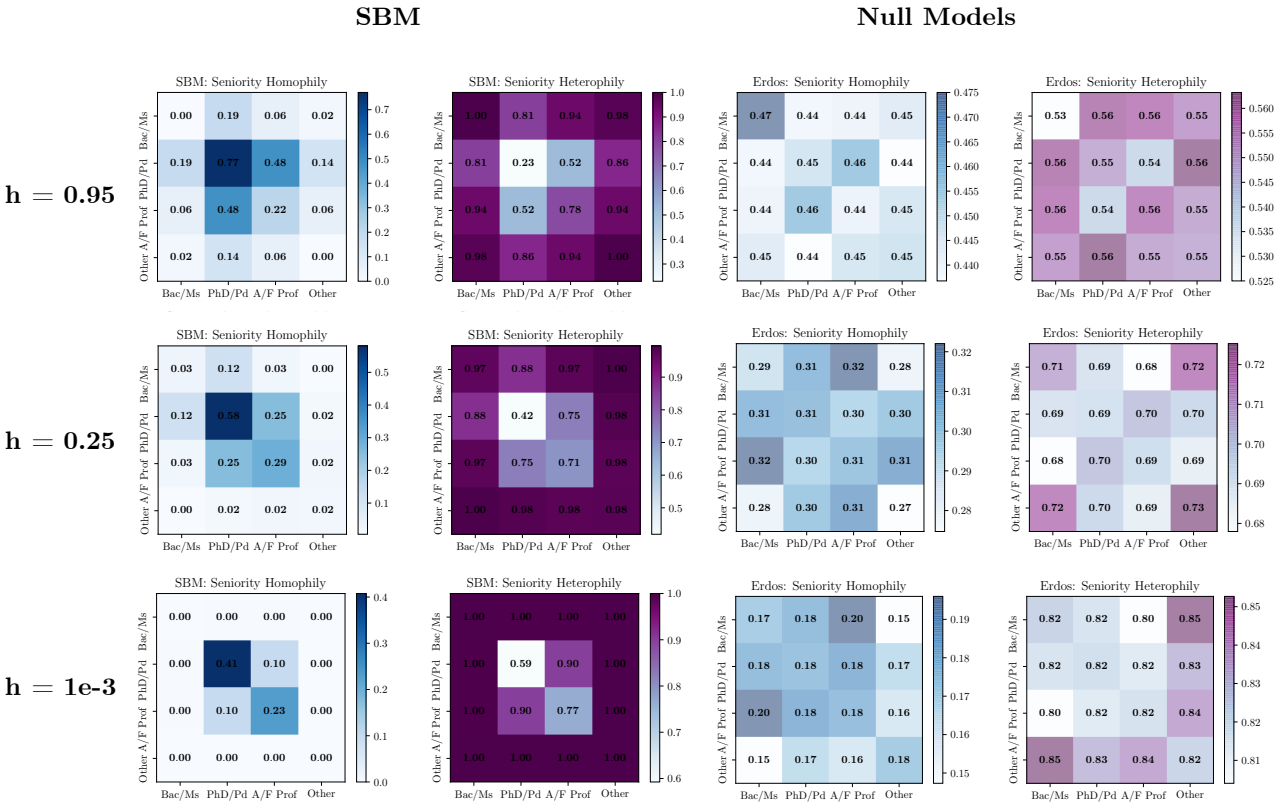


Figure 5: Contact matrices of seniority homophily via. homophily probability within or between groups. Seniority homophily for homophily parameter $h = 0.95, 0.25$ and 0.001 (given same heterophily parameters in age attribute) respectively is represented in both SBM and Null models. Seniority heterophily are computed by the inverse of seniority homophily in contact matrices.

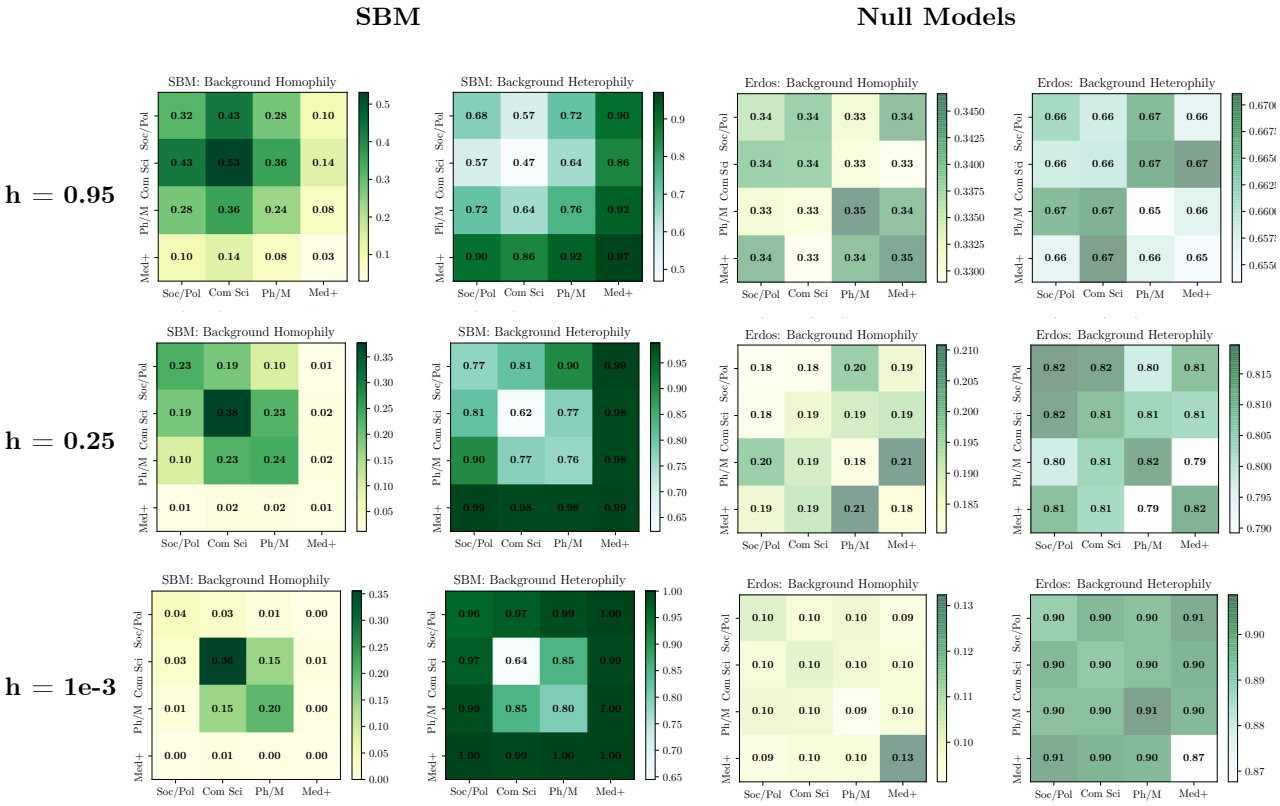


Figure 6: Contact matrices of background homophily via. homophily probability within or between groups. Background homophily for homophily parameter $h = 0.95, 0.25$ & 0.001 (with same set of heterophily parameters in previous attribute) respectively is represented from top to bottom rows. SBM and Null models for background homophily, including heterophily are executed in the similar way as other attributes.

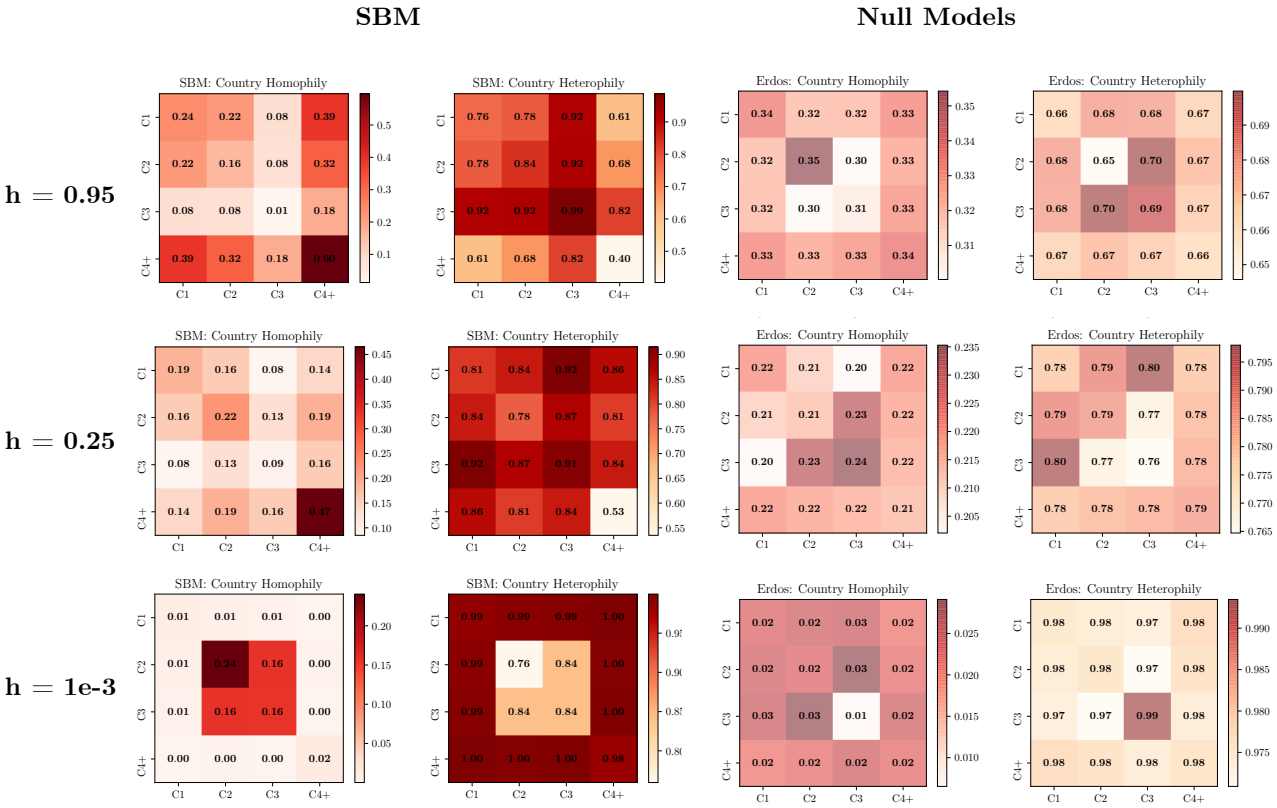


Figure 7: Contact matrices of country homophily via. homophily probability within or between groups. Country homophily for homophily parameter $h = 0.95, 0.25$ & 0.001 (given same set of heterophily parameters) respectively is represented from first to third row of contact matrices. SBM and Null models for country homophily, including heterophily are implemented in the same way as other attributes.

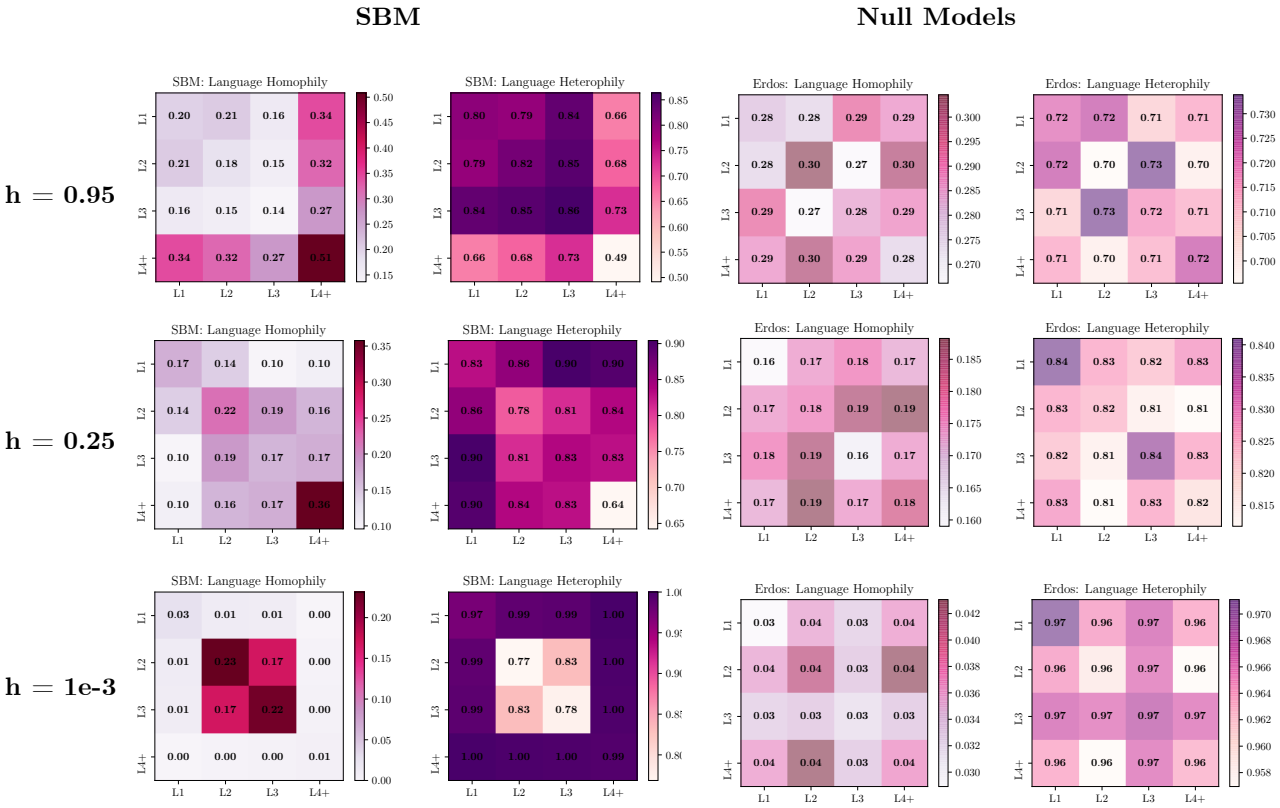


Figure 8: Contact matrices of language homophily via. homophily probability within or between groups. Language homophily for homophily parameter $h = 0.95, 0.25$ & 0.001 (with same set of heterophily parameters) respectively is represented from top to bottom row. SBM and Null models for language homophily, including heterophily are presented in the similar way as other attributes.