# A toy model for topic discovery in social networks and implication on discovered topic diversity

Tuan Pham

## Introduction

As the world becomes more and more connected as well as information increases more and more every day, how do we learn about the currently existing body of knowledge, and at the same time updating with the newly incoming information? Are people becoming more specialized or are there more generalists? Answering these questions might be difficult at this point without assessing simple cases of learning within static networks. Hence I want to examine how different knowledge acquistion strategies could affect one's knowledge set, as well as the diversity of knowledge for the population as a whole.

There could be multiple ways a person could learn something new. For example consider a PhD student reading a paper. In an overly simplified view, the student could start digging down the references for something they have never seen or heard about before and choose to actually read passionately about it, and let's optimistically assume they master it. Further and further down the "rabbit hole", the student starts learning about some of the most arcane, obscure subjects in human's knowledge (**Fig. 1**b).

In another case, the student goes to classes or discusses with their friends about a certain topic, or seeing new papers recommended through the people they follow on social media with the hashtag "#TheNextBigThingIn[*insert-field*]". Through these interactions, they could pick up on something that are new to them and start broadening their horizons, based on their peers' recommendations (**Fig. 1**c).

Considering only these two different ways of learning in a probabilistic sense, I examine the diversity of knowledge, represented as different metrics based on the distribution of topics, as well as graph metrics, in random networks, with and without consideration of modularity within such networks. The results show that the self-learning process tends to improve diversity in a population manner, but the latter process of learning through friends or recommendations would generally benefit individual diversity. Consideration of groups within the models have mixed effects at the individual level more so than the population level.

## Methods

### 1. Model set up

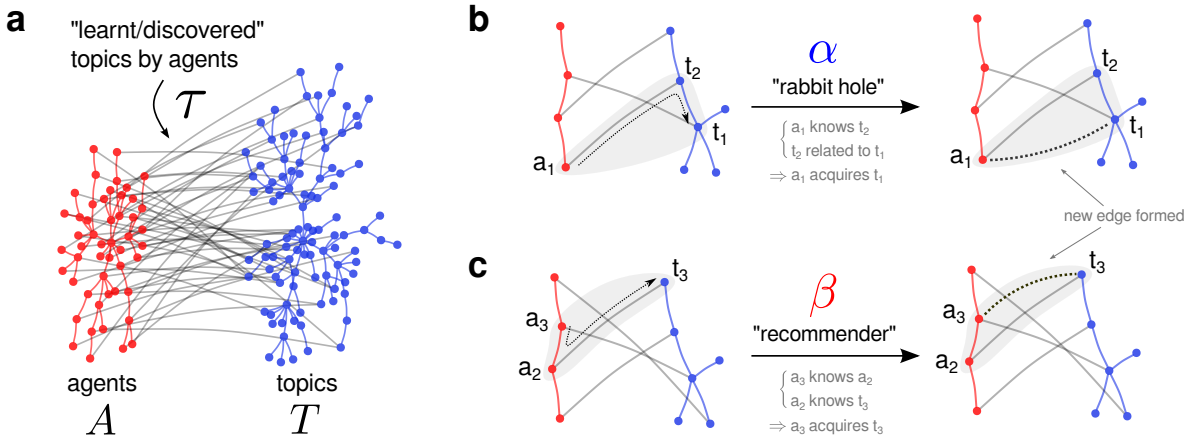


**Figure 1:** *Model setup and description of the update process.* (**a**) Illustration of the intralayer agent graph (red) and topic graph (blue) with the interlayer edges (gray) representing the knowledge set of the agents. (**b**) and (**c**) illustrate the update process either through learning/discovery by related topics ("rabbit-hole") or learning/discovery through friends ("recommender").

**General description:** All models considered here are binary undirected graphs. There are $n_a = 200$ agents and $n_t = 1000$ topics. Denote $A$ and $T$ as the symmetric binary adjacency matrices of the agent graph $G_a$ and topic graph $G_t$ respectively (**Fig. 1**). The bipartite incidence matrix $\tau$ of size $n_t \times n_a$ represents the topics that the agents know about. It is assumed throughout that the intralayer edges are static while the interlayer edges could be "acquired" through the update process. And once an interlayer edge is acquired, it is assumed to be persistent. At the initial stage, each agent is assigned at most $\tau_0 = 5$ topics with certain probabilities based on the models of the intralayer models (see below). There is also an upper limit topic capacity $\tau_{\max} = 50$ per agent, and the update process is only simulated until $1.2\tau_{\max} = 60$ time steps. For each parameter set ($\alpha$, intralayer models, interlayer initialization), I ran 5 simulations each. Hopefully in the future I could acquire more computational resources to run more simulations on larger networks (especially for the

more realistic graphs built from real-world networks).

**Update of interlayer edges:** At each time step, at most one new topic is learnt per agent. The agent could acquire new topic edge either through the "rabbit-hole" strategy with $\alpha$ probability, by learning about the related topics of things an agent already knows about. On the other hand, with probability $\beta$, an agent could acquire a new topic edge by traversing its neighbors in the agent graph then to the topic graph. See **Fig. 2** for illustration of these processes. One way to implement this is below.

Define $\psi(X)$ as a column L1 normalization operation on a matrix $X$, meaning each column vector $\vec{x}_i$ of the matrix is normalized to $\vec{x}_i/||\vec{x}_i||_1$. Define the shorthand notation for the Heaviside function as $[x]_\star = 1$ if $x > 0$, and 0 otherwise.

At each time step, the probability matrix $P$ (of same size as $\tau$) with its column vector $\vec{p}_i$ defining the probability agent $a_i$ choosing a new topic. A way to define this probability is:

$$P = \alpha\psi\left(\left[[T\tau]_\star - \tau\right]_\star\right) + \beta\psi\left(\left[[\tau A]_\star - \tau\right]_\star\right) \tag{1}$$

$$\tau(t+1) \leftarrow \tau(t) + \text{sample}(P) \tag{2}$$

The multiplication steps perform the traversal through neighbors across the intralayer networks. The binarization and subtraction of current $\tau$ simplifies the implementation, to only learn new topics and to balance not being stuck around too popular topics. Additionally, for simplicity here I consider $\beta = 1 - \alpha$ so the process is only defined by $\alpha$. Many other probabilities are ignored as well, for example serendipity (wandering or random discovery of new topics) and forgetting (removal or decrease of strength of interlayer edges). Future implementations should relax these many assumptions for more realistic implementation, and possibly easier to obtain analytical solutions when the nonlinearities are minimized.

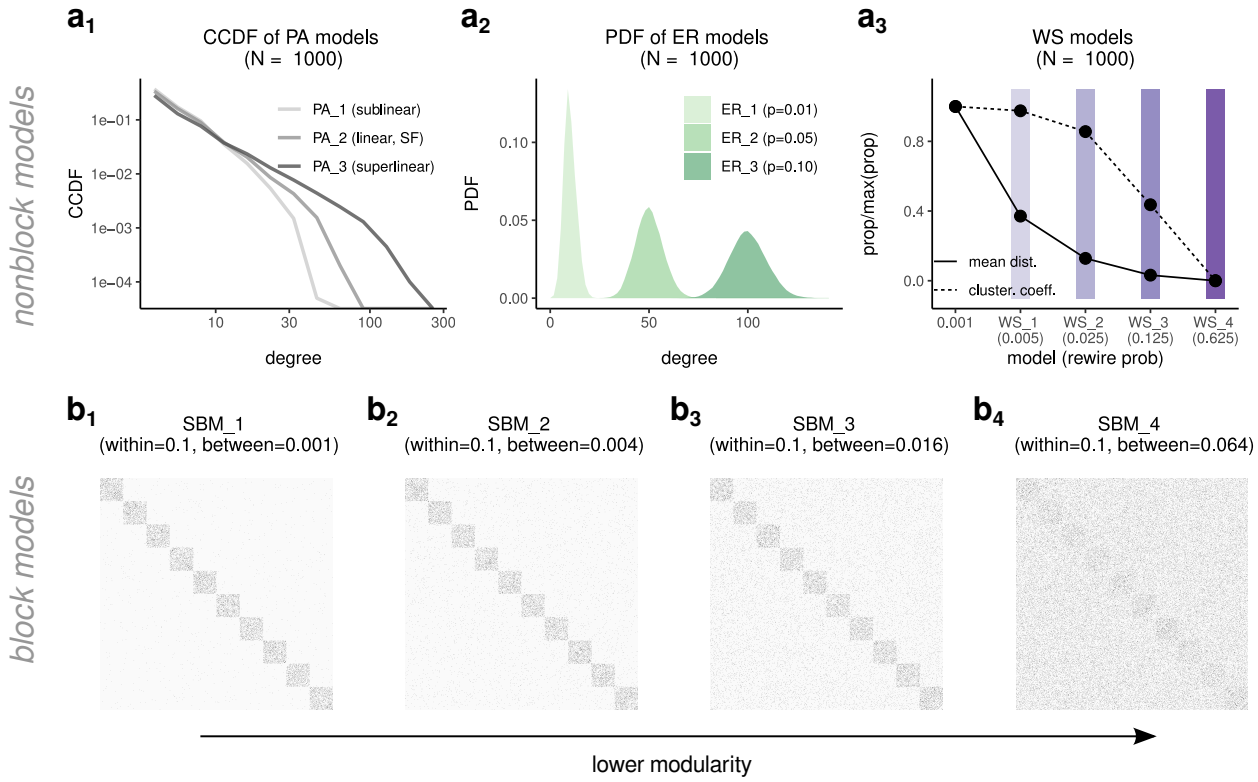


**Figure 2:** *Examples of intralayer static model set up.* For simplicity, for all instances, the types of models used are the same for both agent and topic graph. (**a**) Non-block models (PA: preferential attachment, ER: Erdős–Rényi, WS: Watts–Strogatz; see method for more details). (**b**) Stochastic block models (SBM) with decreasing modularity by increasing the connection probability between members of different groups. There are 10 groups for each intralayer model.

**Intralayer random models:** There are different ways to initialize the intralayer networks to match with different empirical results in real social and knowledge networks. For simplicity, the model types and parameters (except only for the number of nodes) are kept the same for agent and topic graphs during each simulation.

*Nonblock models*: The first approach is non-block networks. These include models constructed from preferential attachment models (PA), in which linear PA would lead to a scale-free network with the scaling parameter $\gamma = 3$. I also take into account nonlinear PA models (see **Fig. 2**$a_1$). I also include comparison with Erdős–Rényi (ER) networks with different connectivity probability (see **Fig. 2**$a_2$), as well as small-world networks generated with the Watts–Strogatz models (see **Fig. 2**$a_3$).

*Block models*: Since in real-world networks, there are usually communities (researchers or papers within the same field),

I use the stochastic block models (SBM) to emulate this with $k_a = k_t = 10$ groups for both agent and topic networks. A way to manipulate these models is to change the probability of connection within groups ($p_{\text{within}}$) or between groups ($p_{\text{between}}$). For simplicity, I kept the former the same while varying the latter (see **Fig. 2**). In retrospect, this makes the models denser by increasing the $p_{\text{between}}$. Future simulations should look for ways to balance this.

Future endeavours should take into account real networks, for example from subsampling a Twitter network as the agent graph and Wikipedia network as the topic graph. Another possibility would be to use citation networks, with authors as agents and papers as topics, groups could be subfields or certain modules in the research topics.

**Interlayer initialization:** Generally at the initialization stage, the probability of connection between a given agent and topic could be the same across topics. However, it is possible that other initialization strategies might bias the results in one way or another. Hence, I introduce two different interlayer initialization strategies, one for *nonblock* intralayer models and one for *block* models. Whenever an initialization method is not mentioned, it is assumed to be the uniform random strategy.

For *nonblock* intralayer models, the probability of connecting to a certain topic could be dependent on its degree in $G_t$. A way to do this is to perform the softmax $\left(\{d_i\};\beta_\sigma\right)$ on the degrees, basically transforming the degree sequence $\{d_i\}$ to a probability distribution. With $\beta_\sigma < 0$ ($\text{SOFTMAX}_1$), low degrees are favored; $\beta_\sigma = 0$ is equivalent to random initialization ($\text{SOFTMAX}_2$), while $\beta_\sigma > 0$ ($\text{SOFTMAX}_3$) favors high degree topics (**Fig. S1**)

For *block* intralayer models, group correspondence could be used as a strategy for initialization as the number of groups are the same for both graphs. This could be parameterized by $p_{\text{sg}}$ (**Fig. S2**) as the probability that agents and topics of the same group ID are connected. The chance $p_{sg} = \frac{1}{k_t} = 0.1$ would be equivalent to random initialization.

### 2. Diversity metric

**Population:** Three population indices are defined. First is $N_T$ the number of distinct topics discovered when taking into account all agents' learnt topics (higher would mean more diverse). Second is the topic population entropy $H_p$, which is the Shannon entropy from the discrete probability distribution of all the topics in the population (higher would mean more diverse). Lastly, taken inspiration from ecological bipartitate network stability analysis, robustness can be calculated, by cumulatively removing random agents and observering the number of distinct topics left. The area under this curve is the robustness (higher would mean a lot of agents are needed to be removed to remove a sufficiently large proportion of topics).

**Individual:** Three individual indices are calculated. $d_g$ is the mean distance of the topics in each agent's learnt topics. In other words, if we define $D(t_i, t_j; G_T)$ as the shortest path distance in $G_t$ between $t_i$ and $t_j$, and an agent $a_k$'s topic set as $\tau(a_k)$ then $d_g(a_k) = E\left[D(t_i, t_j; G_T)\right]_{t_i,t_j\in\tau(a_k)}$. Then I take the mean of these distances across all agents of the agent graph. Higher would mean on average, the agents learn more out of their comfort zone. Another metric is the number connected component $n_{\text{cc}}(a_k)$ of the induced subgraph $G_t(\tau(a_k))$ - higher would mean there are many "islands" of topics that the agent knows about, leaning toward generalist trend. Lastly, the pairwise Jaccard similarity between agents' topic sets are calculated $Js_T$, lower would mean higher local diversity on average. Future endeavours should take into account other metrics for local diversity indices (for example local entropies under subsampling of neighbors or nodes within a certain distance).

**Group:** Additionally, when groups are defined in the block intralayer models, one could also calculate the entropy of the topic group distribution, in both the population sense $H_{\text{gp}}$ and individual sense $H_{\text{gi}}$. More specifically, $H_{\text{gp}}$ is the entropy of the 10 topic groups when taking into account the group identities of all topics learnt by all agents. On the other hand, $H_{\text{gi}}$ is the average entropy of each agent's own topic entropy. These two quantities are different for example, there could be cases where as a population, $H_{\text{gp}}$ is maximized (all groups uniformly distributed) but $H_{\text{gi}}$ could be 0 (each agent only learns about the topics of the same group, leading individual entropy of 0).

## Results

### 1. *Nonblock* intralayer models

The changes of the different diversity metrics for the scale-free networks ($PA_2$) are shown in **Fig. 3** as an example to illustrate the tradeoff effect of the "rabbit-hole" versus "recommender" probability on the population and individual diversity.

Generally, topic population diversity increases with $\alpha$ in terms of the topic entropy $H_g$ and number of topics $N_T$. Through learning/discovery through time, low $\alpha$ could still achive better population diversity. However, it does not seem likely for the worst case considered here, where entropy does not even increase pass its initial value. The initial decrease of $H_g$ when $\alpha = 0.1$ is because the agents start learning from each other, hence temporarily creating bias towards some topics, leading to decrease of entropy. It must be noted here that the entropies are already high initially due to initialization. However, taking the trends of both $N_T$ and $H_g$ into account, it is confident to say that higher $\alpha$ improves topic population diversity. Additionally, higher $\alpha$ leads to more robust retainment of the topics under random agent removal (i.e. higher $R_T$).

On the other hand, topic individual diversity usually decreases based on the chosen metrics. Increased $\alpha$ leads to decreased mean learnt topics distance $d_g$ and number of components $n_{\text{cc}}$ in the induced subgraphs. Intuitively, lower $\alpha$ would allow

H

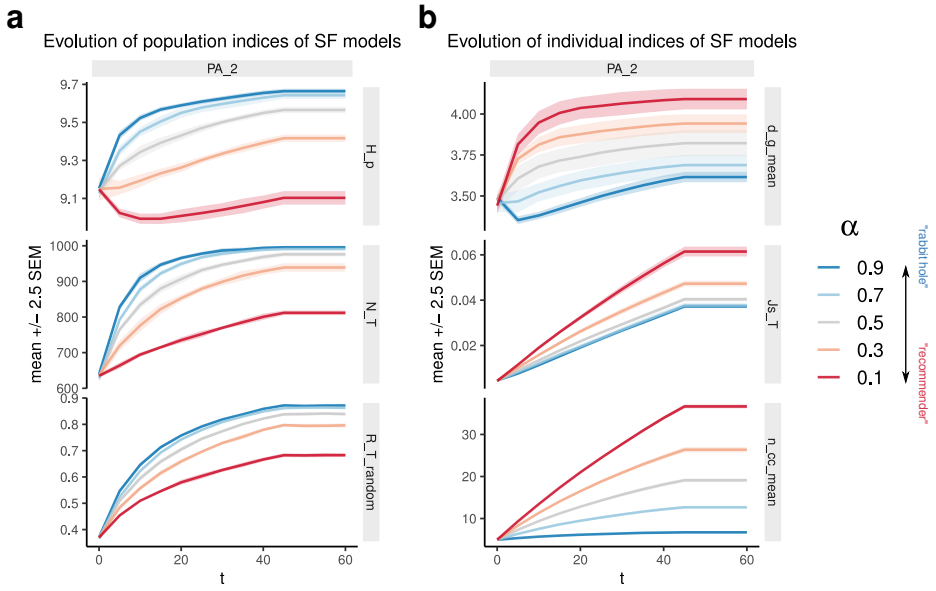


**Figure 3:** *Changes of population diversity indices* (**a**) *and of individual diversity indices* (**b**) *of the scale-free intralayer models* ($PA_2$) *due to* $\alpha$. $H_p$: topic population entropy; $N_T$: number of distinct topics; $R_T$: robustness due to random removal of agents; $d_g$: mean distance of the subset of topics that agents know; $Js_T$: Jaccard similarity of topic set between agents; $n_{cc}$: number of connected components of induced subgraphs based on each agent's learnt topics. See method for detailed description.

the agents to access topics out of their comfort zone easier, hence their own subgraph of topics tend to be more generalist, whereas higher $\alpha$ leads to more specialization. Lastly, at the local level $Js_T$, lower $\alpha$ leads to more similarity between neighbors, hence lower local diversity.

These trends are quite consistent across different considerations of non-block models (**Fig. 4**). Increasing in $\alpha$ leads to higher topic population diversity ($N_T, H_P$), robustness ($R_T$) and local diversity $Js_T$. On the other hand, such increases tend to result in loss of topic individual diversity ($d_g$, $n_cc$). When taking into account degree-dependent initialization strategies (**Fig. S1**), favoring more obscure topics leads to the same trend as random initialization, though small effects when $\alpha$ is small (first 2 rows in **Fig. S3**). However, initially favoring more popular topics actually would be detrimental generally across all population, local and diversity, especially for those networks generated by preferential attachments (PA) models, possibly the learning gets "stuck" around the nodes around those population ones (last row in **Fig. S3**)

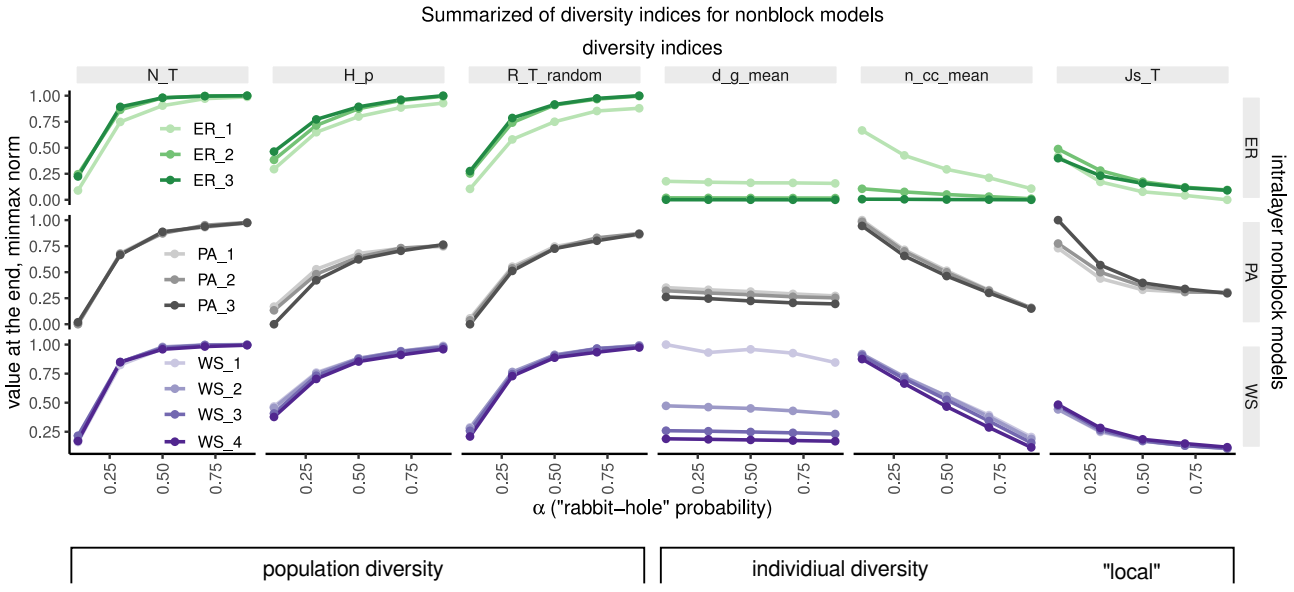


**Figure 4:** *Summary of population and individual diversity indices due to* $\alpha$*, across different nonblock models (with random initialization).* The values here are at the end of the simulation, and min-max normalized within each metric (panels from left to right, see text and **Fig. 3** for description of these different definitions). The different models are shown by colors and panels from top to bottom (see also **Fig. 2** for correspondence of colors).

In summary, in non-block intralayer models, higher $\alpha$ leads to higher topic diversity in a local and population context, but lower individual topic diversity. Initialization favoring more popular topics seems to have a negative effect on these

different metrics.

### 2. *Block* intralayer models and topic group diversity

As real-world networks usually contain communities within them, I use the stochastic block models (SBM) to observe how diversity indices change due to $\alpha$ and network modularity. Generally the trends for population diversity and robustness during the simulation are similar from previously discussed (**Fig. S4**). The trends as a function of model modularity do not seem to differ much either (however the final values do show some differences, and will be discussed later). Looking at the group population entropy $H_{\text{gp}}$ (**Fig. 5**a), only when the networks are less modular do such values show a difference, albeit very small. In particular, only when $\alpha$ is very low that such disadvantage could be visibly inspected (very low $\alpha$ basically leads to learning in groups).

In the individual perspective (**Fig. 5**b), group modularity actually helps with diversity indices $d_g$ and $n_{\text{cc}}$, possibly because there are few long-range links. The trends for local diversity are roughly similar and not affected much by group modularity. However, their final values do show some difference (will be discussed later). Additionally, instead of only looking at topic group diversity in the population sense, one could also inspect it in the individual perspective. On average (panel a, top), for more modular intralayer networks, lower $\alpha$ benefits topic group diversity in the agents, because the agents would have higher chance to learn out of their own comfort zone, especially if their initial topics belong to the same groups. With decreasing group modularity, these differences between $\alpha$ do not seem to matter any more.

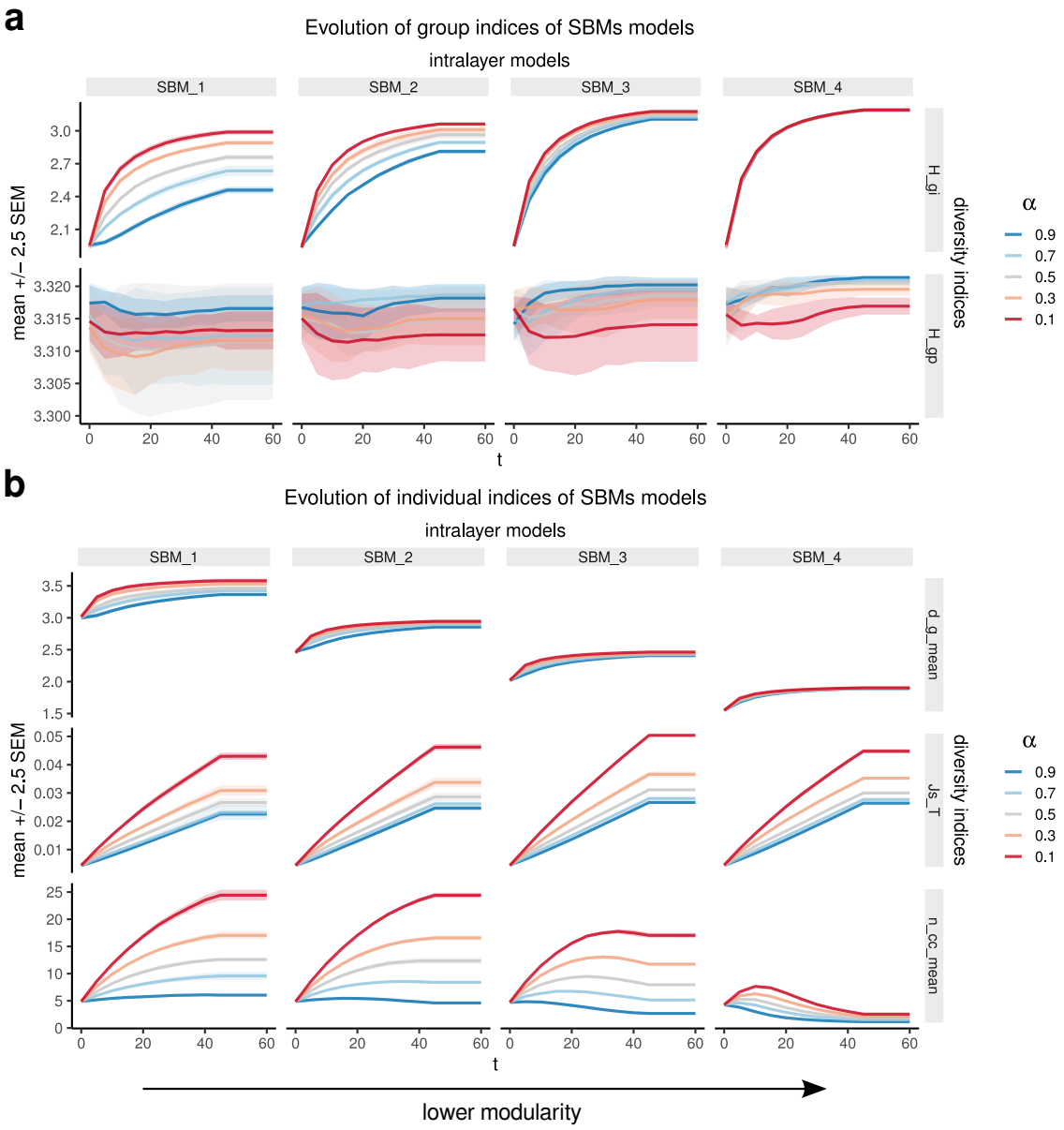


**Figure 5:** *Changes of group diversity indices* (**a**) *and of individual diversity indices* (**b**) *of the stochastic block intralayer models due to* $\alpha$. $H_{\text{gi}}$: topic individual entropy; $H_{\text{gp}}$: topic population entropy. From left to right, the models are designed to have decreasing modularity (see **Fig. 2**). See also **Fig. 3** and text for descriptions of individual diversity metrics.

Inspecting the end values of these different metrics in **Fig. 6** taken into account group-correspondence initialization strategies reveal these effects more clearly.

More specifically, higher $\alpha$ and lower intralayer modularity leads to high population topic diversity and robustness $(N_T, H_p, R_T)$, whereas group correspondence initialization does not seem to have pronounced effects. Group modularity benefits the individual diversity $(d_g, n_{cc})$ but high initial group-correspondence would counter such effects. For local level, group-correspondence does not seem to affect $Js_T$ visibly. However, generally higher $\alpha$ and higher model modularity tends to decrease topic similarity, hence increasing local diversity.

When we start to consider group entropies, generally low group modularity increases both topic group population $(H_{\text{gp}})$ and individual $(H_{\text{gi}})$. High initial correspondence though seems to benefit group population diversity (although such benefits may be small, see **Fig. 5**), it seems to decreases group individual diversity.

In summary, with consideration of intralayer block models, higher $\alpha$ still benefits more for population diversity and robustness, but not so much with group population diversity. High network modularity may hurt population diversity, regardless of initial group-correspondence. On the other hand, lower $\alpha$ is more beneficial for individual indices, including group individual diversity, but either low network modularity or higher group correspondence would tend to be harmful for these metrics. At the local level, higher $\alpha$ and high network modularity tends to be more beneficial for decreasing similarity between agents.

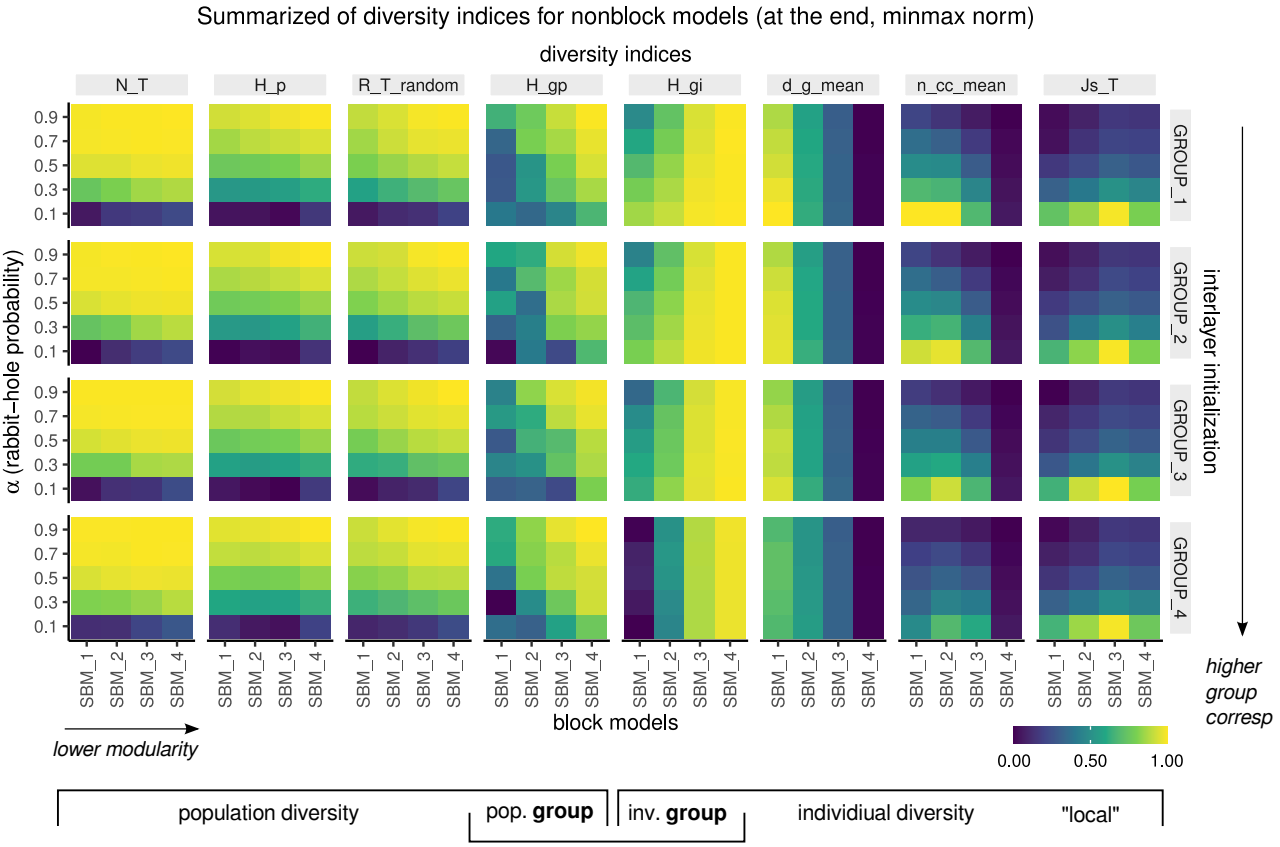


**Figure 6:** *Summary of population and individual diversity indices due to $\alpha$, across different block models.* Within each panel, x-axis shows decreasing modularity of intralayer model, y-axis is $\alpha$ while the color represents values here at the end of the simulation, and min-max normalized within each metric. From left to right are different diversity metrics (see text for more information). From top to bottom are different group correspondence initialization strategies (top is equivalent to a random initialization; see **Fig. S2** for example)

## Discussion

In conclusion, with this simple toy model of topic discovery and a simple update rule depending on the probability of traversing through neighbors of bipartite networks, some interesting results are obtained. First of all, increasing in $\alpha$ (self-learning, traversing through interlayer edges first) leads to higher topic population diversity and robustness in various random models for the intralayer networks, including blocks and non-block models. However, such increase has drawbacks when looking at topic individual diversity, as it reduces the chance for the agent nodes to acquire interlayer edges from topics that are usually "out-of-their-comfort-zone". Traversing through intralayer edges first ($\beta$ route) would better benefit individual diversity.

When groups are considered in the intralayer networks, group modularity may hurt the population diversity (though some only by a little) and more apprently for group individual entropy, though interestingly more beneficial for individual indices through the lense of graph distances and components. Although initial group correspondence does not have much effects on the population diversity, it has a dramatic drawback at the individual level (both entropy and graph metrics).

Though there are interesting results in a theoretical sense as a toy model, there are quite many limitations of the current model. Future studies should relax the assumptions made here and test out different versions of the models, for example inclusion of directed weighted edges (strengths could imply confidence in knowledge in $\tau$), non-persistent interlayer edges, different update probabilities (serendipity, forgetting, strengthing, ...), the cost of learning new subjects, delays in acquiring new knoweldge, different versions of the update equation. Furthermore, future endeavours should take into account performing the update process in real networks, which could be constructed using, as an example, the citation networks (agents as authors, papers as topics, groups as fields or subfields).

Additionally, even within the simulations already generated by this project, there are other directions for further explorations, including examination of the modularity changes in the bipartite $\tau$ or in the projected topic graphs (for example, low $\beta$ might start to create communities as evidenced by high Jaccard similarity), the distribution of specificialists and generalists, as well as calculation of nestedness (due to the defined maximum capacity $\tau_{\max}$, the density of the networks would already be controlled). Furthermore, due to the assumption of persistent interlayer edges, it is possible to characterize properties of persistent homlogy of the networks (or the projected version).

In a bigger picture, regardless of questions about information or diversity, how might the process laid out by this model be relevant to other systems? Due to the simplicity set out by the model and the update process, there could be some biological interpretations, though quite loosely and to be taken lightly. For example, in an ecological perspective, one could consider the agent graph as species with links as certain phenotypic/genotypic similarity, and the topic graph as the different environments with links as environmental similarity. The interlayer edges would represent adaptation of a given species to a given environment. The $\alpha$ route would, in a sense, be analogous to adaptation to new environments that share certain similarity with already-adaptated environments. On the other hand, the $\beta$ route could be adaptation to a new environment that the animal has never been in, yet in nature there already exists similar species that has adaptated to such environment, hence such new adaptation is probable.

Another example can be from the neuroscience perspective, albeit in an oversimplified fashion and possibly biologically implausible. However, as a thought model, the agent graph could be neurons with links between them as correlations between them (either structurally via synaptic connections or functionally through activity). The topic graph could be inputs, in which the intralayer edges could represent either coincidence between these different types of inputs or similarity between them. The interlayer edges could represent which inputs a given neuron could code for. The update process shares some inspiration from associative learning. The $\alpha$ route may represent neurons acquire new inputs to code for due to co-occurence of inputs or just because the inputs are sufficiently similar. The $\beta$ route could be because the neuron gets recruited into the engram encoded for a certain object or abstract concept due to the correlative activity with another neuron that already codes for such object.

## Acknowledgement

I would like to thank Dr. Mercedes Pascual and Dr. Sergio A. Alcala Corona, along with every one in the Network of Ecology and Evolution class, and my friend Poojya Ravishankar for the discussion and feedbacks on this toy model.

## Supporting information

### Code and data availability

Current code repository lies here: `https://github.com/tuanpham96/topic-diversity` (some updates might be a bit slow push up as I'm in the middle or reorganization). The data are simulations (hence quite large), so I will not be able to make them available straight on Github. However, I'm happy to share them.

## Supplementary figures

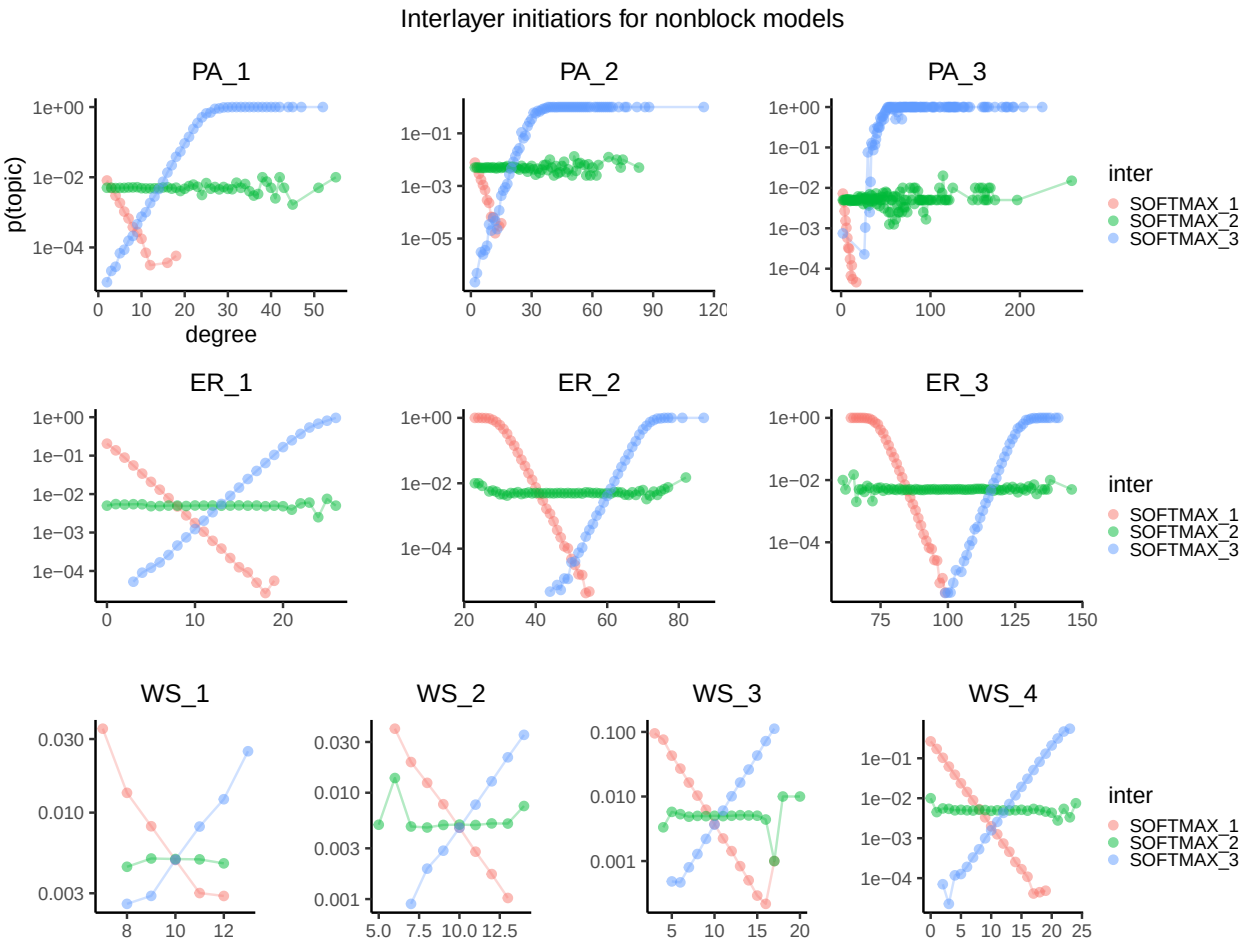


**Figure S1:** *Different initialization strategies based on the degree of the nodes within each nonblock random networks (favor low degree; random; fave high degree).* Note that the probabilities here are plotted in log scale, against the degree of nodes.

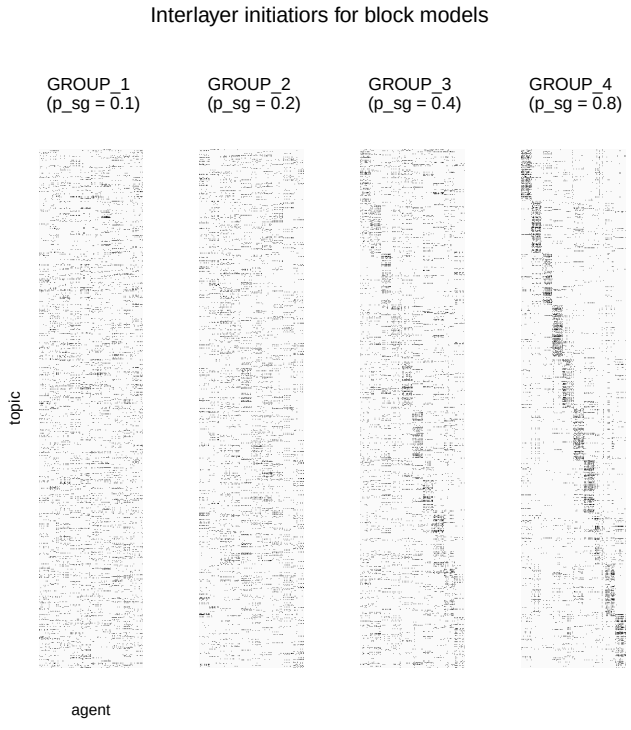


**Figure S2:** *Different group correspondence initialization strategies for intralayer block models*

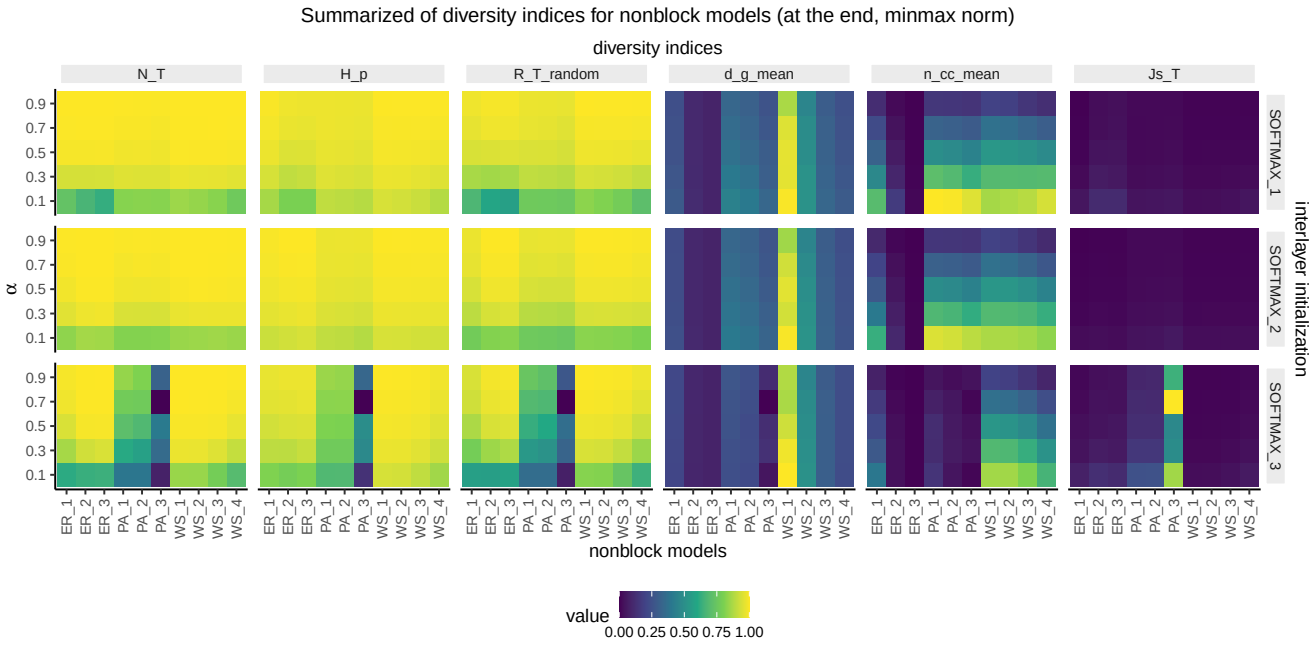


**Figure S3:** *Summary of population and individual diversity indices due to $\alpha$, across different nonblock models and initialization strategies (see **Fig. S1**)*

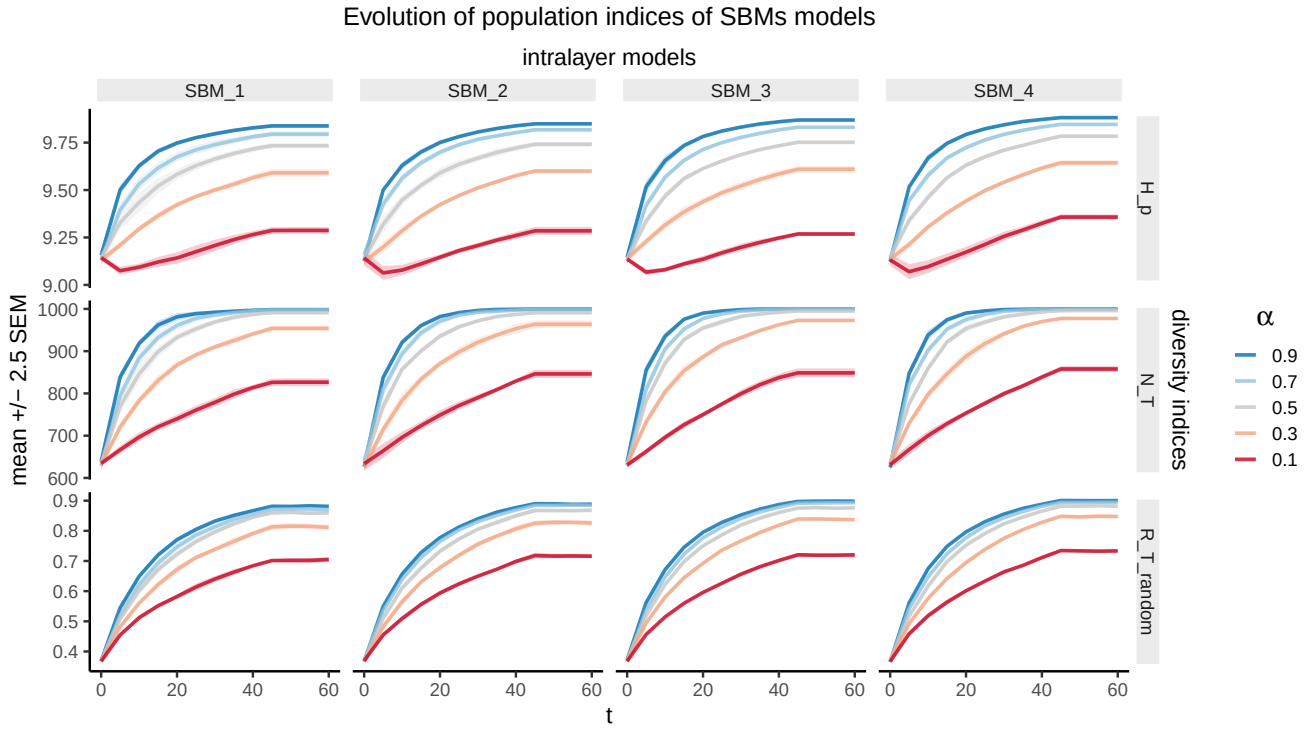


**Figure S4:** *Changes of population diversity indices of the stochastic block intralayermodels due to $\alpha$*