Modelling the effects of self-learning and social influence on the diversity of knowledge

Tuan Pham

In this paper, I present a computational model of acquiring new knowledge through self-learning (e.g. a Wikipedia “rabbit hole”) or social influence (e.g. recommendations of friends). The model is a bipartite network between a static social network (*agents*) and a static knowledge network (*topics*). For simplicity, the learning process is singly parameterized by as the probability of self-learning, leaving as the socially-influenced discovery probability. Numerical simulations show a tradeoff between on the diversity of knowledge when examined at the population level (e.g. number of distinct topics) and at the individual level (e.g. the average distance between topics for an agent), consistent across different intralayer configurations. In particular, higher values of , or increased self-learning, lead to higher population diversity and robustness. However, lower values of , where learning is influenced by to social inputs, expand individual knowledge diversity more readily. These numerical results provide basic insight into how social influences can affect the diversity of human knowledge, especially in the age of information and social media.

# 1. Introduction

As the world becomes more connected and the amount of information in it increases, how do we learn an existing body of knowledge, and simultaneously combine that with new incoming information? How diverse is knowledge acquired through social interactions? Do people become more specialized or generalized in the face of of new information and limited cognitive capacity? Increased specialization, in particular, has implications for creativity and research productivity [[1]](#bookmark=id.49x2ik5).

Answering these questions is difficult without assessing simple cases of learning within static networks. Previous work in the field has addressed topic diversity analysis at the population level without considering the evolution of new topic acquisition [[2]](#bookmark=id.2p2csry), or used dynamic processes and analyses within only the intralayer networks [[3]](#bookmark=id.147n2zr). A recent work [[4]](#bookmark=id.3o7alnk) does address the involvement of social interaction in innovation dynamics, but does not directly address the level of social influence or the resulting knowledge diversity. Thus, I examined how different knowledge acquisition strategies could affect individual knowledge sets as well as the diversity of knowledge for the whole population.

There are multiple ways a person can learn something new. Here I focus on two ways: (1) active *self-learning* by acquiring new knowledge through related topics (**Fig.** [**1**](#bookmark=id.3fwokq0)b); and (2) through *social influence* as suggested by one’s own social circle (**Fig.** [**1**](#bookmark=id.3fwokq0)c). An example of the former is following a “rabbit hole”, starting from an already-known topic on Wikipedia or a reference in a bibliography. On the other hand, examples of the second scenario include movie recommendations from friends or new sources shared via social media. As humans tend to have limited capacity for learning, I ask how these two scenarios affect the diversity of knowledge of different *individuals* on average (specialists versus generalists), and of the entire social network *population* as a whole (over all topics).

Considering only these two different ways of acquiring new topics in a probabilistic manner, I examine the diversity of knowledge, represented as different metrics based on the distribution of topics, as well as graph metrics. These are examined in simulations of randomly generated networks, with and without consideration of modularity within such networks. The results show that the self-learning processes tend to improve diversity in the population, but recommendations through social influence generally benefit individual diversity. Consideration of groups within the models have mixed effects at the individual level more than the population level.

# 2. Methods

## 2.1. Model

### General description

All models considered here are binary undirected graphs. There are agents and topics. Denote and as the symmetric binary adjacency matrices of the agent graph and topic graph respectively (**Fig.** [**1**](#bookmark=id.3fwokq0)a). The bipartite incidence matrix of size 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 topics with certain probabilities based on the models of the intralayer models (see below). There is also an upper limit topic capacity per agent, and the update process is only simulated until time steps. For each parameter set (, intralayer models, interlayer initialization), I ran 5 simulations each.

### Update of interlayer edges

At each time step, at most one new topic is learnt per agent. The agent could acquire a new topic edge either through the self-learning strategy with probability, by learning about the related topics of things an agent already knows about (**Fig.** [**1**](#bookmark=id.3fwokq0)b). On the other hand, with probability , an agent could acquire a new topic edge by traversing its neighbors in the agent graph then to the topic graph (**Fig.** [**1**](#bookmark=id.3fwokq0)c). One way to implement this is below.

Define as a column L1 normalization operation on a matrix , i.e. each column vector of the matrix is normalized to . Define the shorthand notation for the Heaviside function as if , and otherwise. At each time step, the probability matrix (of same size as ) with its column vector defining the probability agent choosing a new topic. A way to define this probability is:

The multiplication steps traverses through neighbors across the intralayer networks. The binarization and subtraction with the current simplifies the implementation, where the agent only learns new topics and avoids already-popular topics. Additionally, for simplicity here I consider so the process is only defined by . 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).

### Intralayer random models

For simplicity, the model types and model hyper-parameters (except only for the number of nodes) are similar the same for agent and topic graphs for each simulation.

*Nonblock models*: The first approach is non-block networks. In the main text, I analyze the model scale-free network (SF), constructed by the linear from preferential attachment models (PA) [[5]](#bookmark=id.23ckvvd) (see **Fig.** [**2**](#bookmark=id.4f1mdlm)). Additionally, I analyze nonlinear PA models, Erdős–Rényi (ER) networks [[6]](#bookmark=id.ihv636) with different connectivity probability, as well as small-world networks generated with the Watts–Strogatz (WS) models [[7]](#bookmark=id.32hioqz) (see **Fig. S1**a).

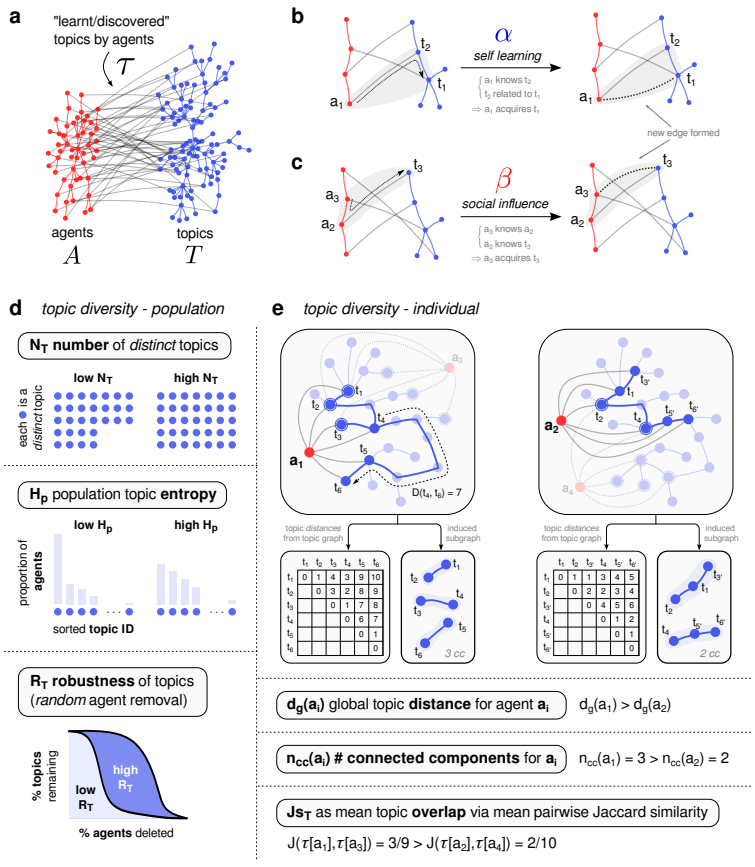
*Block models*: Since in real-world networks, there are usually communities (researchers or papers within the same field), I use the stochastic block models (SBM) [[8]](#bookmark=id.1hmsyys) to emulate this with groups for both agent and topic networks. A way to manipulate these models is to change the probability of connection within groups () or between groups (). For simplicity, I kept the former the same while varying the latter (see **Fig.** [**3**](#bookmark=id.19c6y18)).

### Interlayer initialization

At the initialization stage, the probability of connection between a given agent and topic is uniform across topics, but other possible initialization strategies could influence results. 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 depends on its degree in . A way to do this is to perform the on the degrees, basically transforming the degree sequence to a probability distribution. With (), low degrees are favored; is equivalent to random initialization (), while () favors high degree topics (**Fig. S2**a)

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 (**Fig.** [**3**](#bookmark=id.19c6y18)) as the probability that agents and topics of the same group ID are connected. The chance would be equivalent to random initialization.



**Figure 1** *Description of the topic update/discovery process in the model and the different knowledge diversity metrics.* (**a**) Illustration of the intralayer agent graph (red) and topic graph (blue) with the interlayer edges (gray) representing the knowledge set of the agents. Gray triangles in (**b**) and (**c**) illustrate the update process either through learning/discovery by related topics (self-learning) or learning/discovery through friends (social influence). (**d**) Illustrations of different diversity metrics at the population level (each blue circle is a topic). (**e**) Illustrations of topic diversity metrics at the individual and local level (see **Sect.** [**2.2**](#bookmark=id.1v1yuxt) for detailed descriptions).

## 2.2. Diversity metric

These diversity metrics are illustrated in **Fig.** [**1**](#bookmark=id.3fwokq0)d,e.

### Population

Three population indices are defined, taken from ecological perspective [[9]](#bookmark=id.41mghml). First, is the number of distinct topics discovered when taking into account all agents’ learned topics, where higher values correspond to more diversity. Second, is the topic population entropy – the Shannon entropy from the discrete probability distribution of all the topics in the population; again, higher values correspond to more diversity. Lastly, taking inspiration from ecological network stability analysis [[10]](#bookmark=id.2grqrue), robustness can be calculated by cumulatively removing random agents and observing the remaining percentage of distinct topics. The area under this curve is the robustness , where higher indicate that many agents must be removed to remove a large proportion of topics.

### Individual

Three individual indices are calculated and the averaged computations across nodes (or pairs) of the agent graph are reported. First, is the mean distance of the topics in each agent’s learned topics. In other words, if we define as the shortest path distance in between and , and an agent ’s topic set as then . Higher values suggest that the agents know more topics and tend more towards generalists. Another metric is the number-connected component of the induced subgraph , where higher values mean that there are many “islands” of topics that the agent knows, again leaning toward generalist. Lastly, the mean pairwise Jaccard similarity between agents’ topic sets are calculated, lower values indicate higher local diversity.

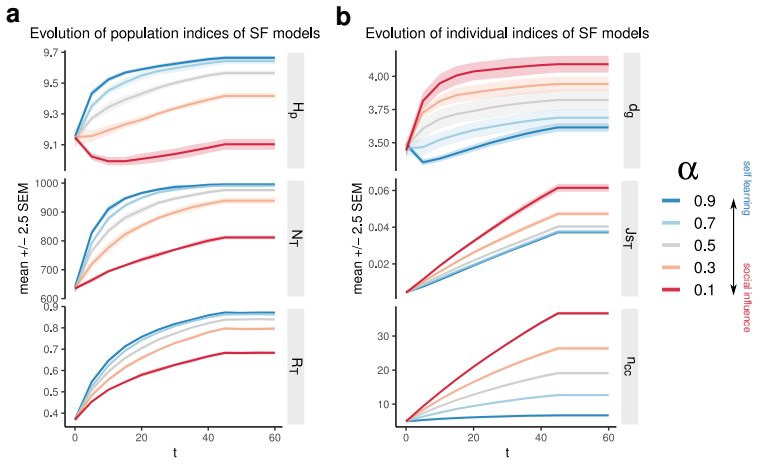
### Group

When groups are defined in the block intralayer models, I calculated the entropy of the topic group distribution, in both the population sense and individual sense . More specifically, 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, is the average entropy of each agent’s own topic entropy. These two quantities are different; e.g. can be maximized (all groups uniformly distributed) while is 0 (each agent only learns about the topics within the same group).

### Code availability

The source code is at <https://github.com/tuanpham96/topic-diversity>. The simulations were run in parallel on Azure VM.

# 3. Results



**Figure 2** *Changes of population topic diversity indices* (**a**) *and of individual diversity indices* (**b**) *of the scale-free (SF) intralayer models due to .* : topic population entropy; : number of distinct topics; : robustness due to random removal of agents; : mean distance of the subset of topics that agents know; : Jaccard similarity of topic set between agents; : number of connected components of induced subgraphs based on each agent’s learned topics. See **Sect.** [**2.2**](#bookmark=id.1v1yuxt) and **Fig.** [**1**](#bookmark=id.3fwokq0)d,e for more details. Each line represents the mean changes of 5 realizations, analyzed every 5 steps.

## 3.1. *Nonblock* intralayer models

The changes of the different diversity metrics for the scale-free networks are shown in **Fig.** [**2**](#bookmark=id.4f1mdlm) as an example to illustrate the tradeoff effect of the self-learning versus social influence probability on the population and individual diversity.

Generally, topic population diversity increases with self-learning probability in terms of the topic entropy and number of topics . Through learning/discovery through time, low could still achieve better population diversity. However, it does not seem likely for the worst case considered here, where entropy does not increase from its initial value. The initial decrease of when 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 and into account, we see that higher improves topic population diversity. Additionally, higher leads to more robust retainment of the topics under random agent removal (i.e. higher ).

On the other hand, topic individual diversity decreases based on the chosen metrics. Increased leads to decreased mean learned-topics distance and number of components in the induced subgraphs. Intuitively, higher social influence – lower – would allow the agents to access topics outside of their comfort zone more easily, hence their own subgraph of topics tend to be more generalist, whereas higher leads to more specialization. Lastly, at the local level , lower leads to more similarity between neighbors, hence lower local diversity. Although not analyzed, this hints at how social influence could create modularity in the learnt topic graph .

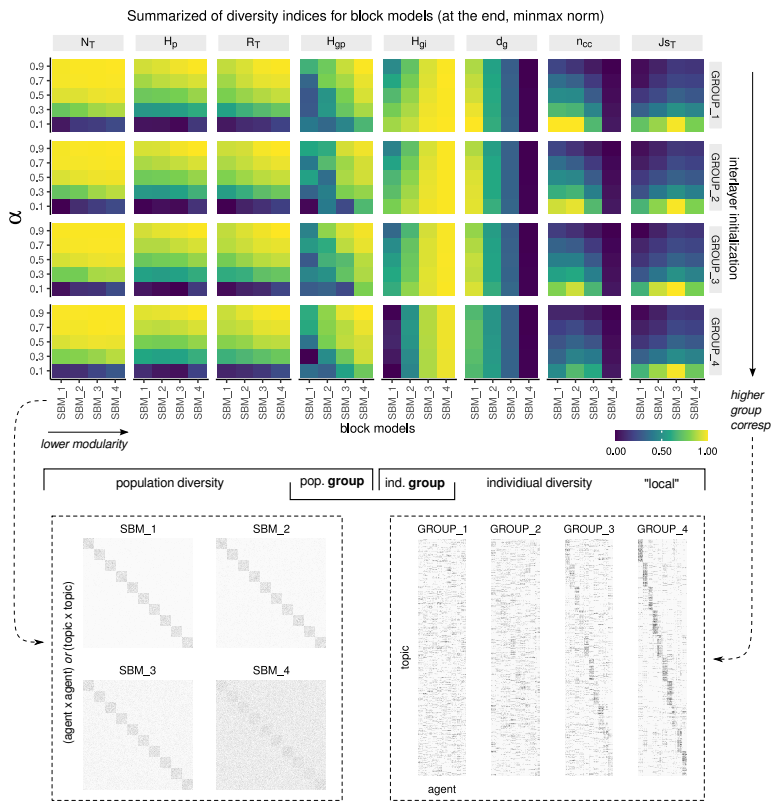
These trends are consistent across different considerations of non-block models (**Fig. S1**). Increasing leads to higher topic population diversity (), robustness () and local diversity (). On the other hand, those increases result in loss of topic individual diversity (, ). When taking into account degree-dependent initialization strategies (**Fig. S2**), favoring more obscure topics leads to the same trend as random initialization. However, initially favoring more popular topics is detrimental across population, local and individual diversity indices, especially for networks generated by preferential attachments (PA) models, possibly because learning more easily gets stuck in topics connected to the popular ones (last row in **Fig. S2**b).

In summary, in non-block intralayer models, higher values of self-learning lead to higher topic diversity in a local and population context, but higher values of social influence encourage individual topic diversity. Initializations that favor more-popular topics have a negative effect on these metrics.

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

As real-world networks usually contain communities within them, I use the stochastic block intralayer models (SBM) to observe how diversity indices change due to and network modularity. Generally, the trends for population diversity and robustness during the simulation are similar to those from previously discussed (**Fig. S3**a). The trends as a function of model modularity also do not differ. Looking at the group population entropy (**Fig. S3**b, bottom), only when the networks are less modular do those values show a difference, albeit very small.

In the individual perspective (**Fig. S3**c), I note that group modularity increases diversity indices and , possibly because there are fewer long-range links. The trends for local diversity are similar and not affected by group modularity. Additionally, instead of only looking at topic group diversity in the population sense, one could also inspect it in the individual perspective. On average (**Fig. S3**b, top), for more modular intralayer networks, social influence benefits topic group diversity in the agents, because the agents would have more chances 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 do not seem to matter.



**Figure 3** *Summary of population and individual diversity indices due to , across different block models.* Within each heatmap, x-axis shows decreasing modularity of intralayer model (via increasing inter-modular connectivity), y-axis is . The color represents values at the end of the simulations, and min-max normalized within each metric. From left to right are different diversity metrics. From top to bottom are different group correspondence initialization strategies.

Inspecting the end values of these different metrics in **Fig.** [**3**](#bookmark=id.19c6y18) taken into account group-correspondence initialization strategies reveal these effects more clearly.

More specifically, higher and lower intralayer modularity generally leads to higher population topic diversity and robustness (), whereas group correspondence initialization does not seem to have pronounced effects. Group modularity benefits individual diversity () but high initial group-correspondence would counter such effects, as learning gets stuck within communities. At the local level, group-correspondence does not seem to affect visibly. However, generally higher and higher model modularity tends to decrease topic similarity, hence increasing local diversity.

For group entropies, low group modularity increases both topic group population () and individual () diversity. High initial correspondences benefit group population diversity (although such benefits may be small, see **Fig. S3**b), but decreases group individual diversity.

In summary, with consideration of intralayer block models, higher influences population diversity and robustness, but less so the group population diversity. In the presence of high social influence, high network modularity may hurt population diversity, regardless of initial group-correspondence. On the other hand, lower is generally more beneficial for individual indices, like those discussed with SF models, including group individual diversity, but either low network modularity or high group correspondence are harmful for these metrics. At the local level, higher and high network modularity decrease similarity between agents.

# 4. Discussion

In conclusion, with this simple toy model of topic discovery and a simple update rule that depends on the probability of traversing neighbors of bipartite networks, several interesting results emerge. First, increasing (self-learning, traversing through interlayer edges first) leads to higher topic population diversity and robustness in several random models for the intralayer networks, including blocks and non-block models. However, those increases have drawbacks for topic individual diversity, as it reduces the chance for the agent nodes to acquire interlayer edges from topics that are distant from their comfort zone. Social influence, traversing through intralayer edges first ( route) better influences individual diversity.

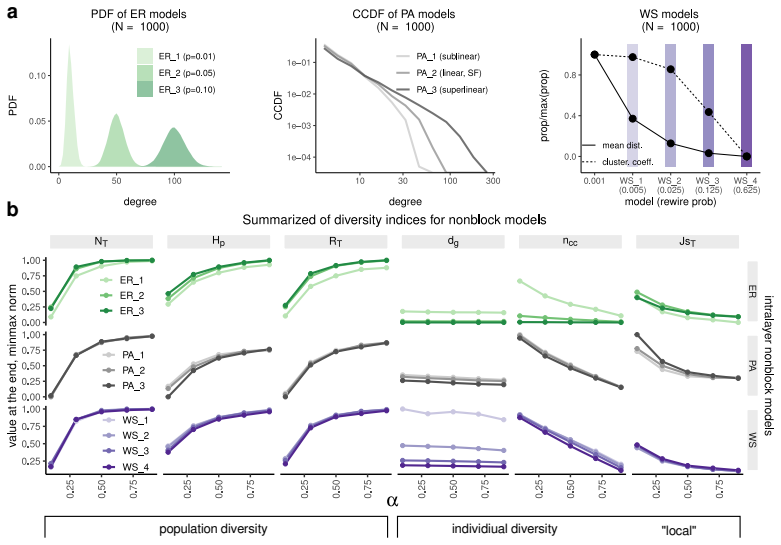
When groups are considered in the intralayer networks, group modularity may hurt the population diversity (some more than others) and more apparently for group individual entropy, though interestingly more beneficial for individual indices through the lens of graph distances and components. Although initial group correspondence does not influence population diversity, it has a dramatic drawback at the individual level (both entropy and graph metrics).

There are limitations of the current model. Future studies should relax the assumptions made here and test out different versions of the models, e.g. different ratios between intralayer network sizes, inclusion of directed weighted edges (strengths could imply confidence in knowledge in ), non-persistent interlayer edges, different update probabilities (serendipity, forgetting, strengthening, ...), the cost of learning new subjects, delays in acquiring new knowledge, different versions of the update equation and, more importantly, the dynamic nature of the intralayer networks (e.g. [[3]](#bookmark=id.147n2zr)) and the decreased disruptiveness in new knowledge discovery [[11]](#bookmark=id.vx1227). 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) or social networks [[2]](#bookmark=id.2p2csry). Additionally, further analyses should include examination of the modularity changes in the bipartite [[12]](#bookmark=id.nmf14n) or in the projected graphs (for example, low might start to create communities as evidenced by high Jaccard similarity in these simulations), the distribution of specialists and generalists, different local diversity definitions (e.g. topic entropy as a function of distance from a given agent) and persistent homology analyses (since the interlayer edges are defined as persistent here).

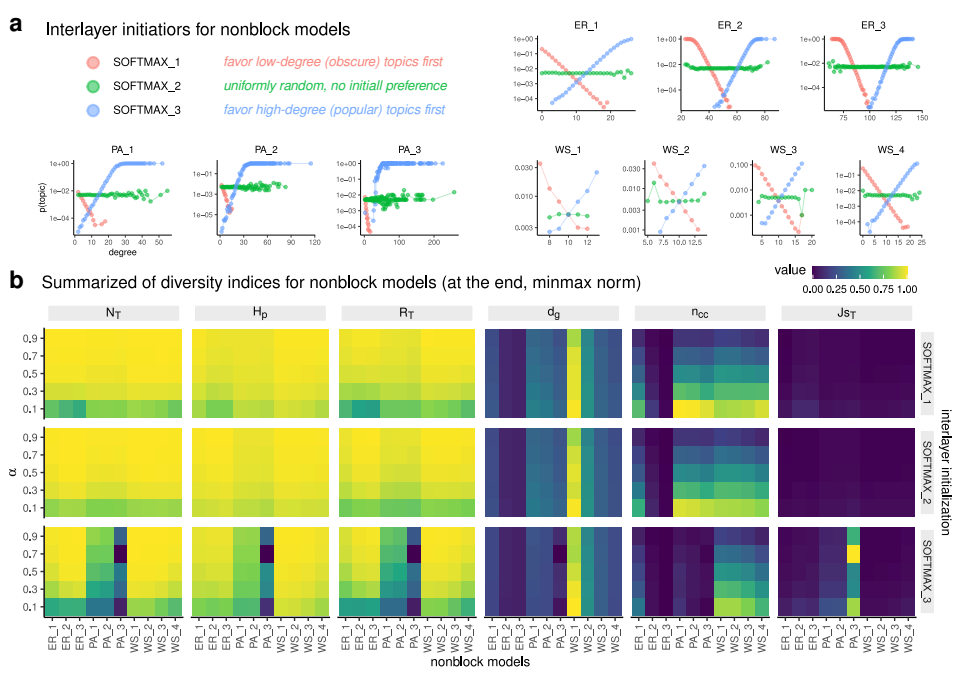
# 5. Acknowledgement

I would like to thank Dr. Mercedes Pascual and Dr. Sergio A. Alcala Corona, my friends Sam Nguyen and Poojya Ravishankar for their discussion and feedbacks of the model and interpretations.

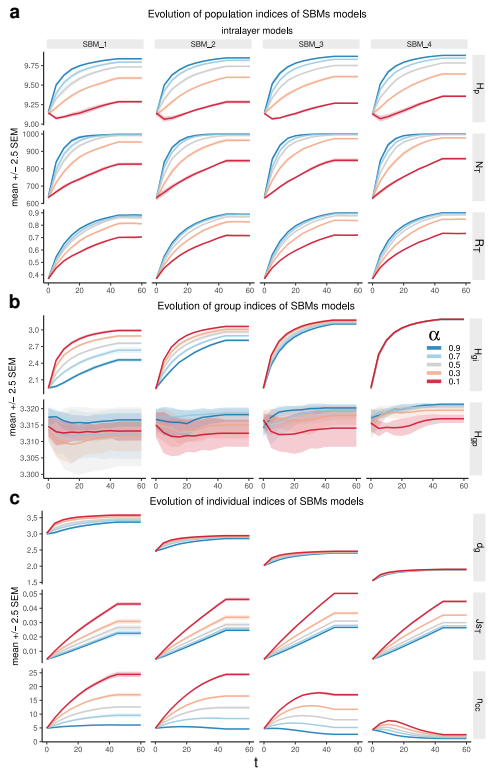
# Supplementary figures



**Figure S1** *Variations of nonblock intralayer models.* (**a**) Set up of nonblock models. PA: preferential attachment, ER: Erdős–Rényi, WS: Watts–Strogatz (**Sect.** [**2.1**](#bookmark=id.37m2jsg)) (**b**) Changes of diversity indices for these models as a function (**Sect.** [**2.2**](#bookmark=id.1v1yuxt))



**Figure S2** *Different initialization strategies for nonblock models* (**a**) *based on the topic intralayer degrees and* (**b**) *effects on population and individual diversity indices as a function of .* See **Fig. S1**a for names and illustrations of the different models.



**Figure S3** *Changes of population diversity indices* (**a**)*, group diversity indices* (**b**) *and individual diversity indices (****c****) for the stochastic block intralayer models due to .*

# References

[1] F. Teodoridis, M. Bikard, and K. Vakili, “Creativity at the knowledge frontier: The impact of specialization in fast- and slow-paced domains,” *Adm. Sci. Q.*, vol. 64, no. 4, pp. 894–927, Dec. 2019, doi: [10.1177/0001839218793384](https://doi.org/10.1177/0001839218793384).

[2] L. Weng and F. Menczer, “Topicality and impact in social media: Diverse messages, focused messengers,” *PLoS One*, vol. 10, no. 2, p. e0118410, Feb. 2015, doi: [10.1371/journal.pone.0118410](https://doi.org/10.1371/journal.pone.0118410).

[3] Y. Sun and V. Latora, “The evolution of knowledge within and across fields in modern physics,” *Sci. Rep.*, vol. 10, no. 1, p. 12097, Jul. 2020, doi: [10.1038/s41598-020-68774-w](https://doi.org/10.1038/s41598-020-68774-w).

[4] I. Iacopini, G. Di Bona, E. Ubaldi, V. Loreto, and V. Latora, “Interacting discovery processes on complex networks,” *Phys. Rev. Lett.*, vol. 125, no. 24, p. 248301, Dec. 2020, doi: [10.1103/PhysRevLett.125.248301](https://doi.org/10.1103/PhysRevLett.125.248301).

[5] A. L. Barabasi and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999, doi: [10.1126/science.286.5439.509](https://doi.org/10.1126/science.286.5439.509).

[6] P. Erdös and A. Rényi, “On random graphs I,” *Publicationes Mathematicae*, vol. 6, pp. 290–297, 1959.

[7] D. J. Watts and S. H. Strogatz, “Collective dynamics of ’small-world’ networks,” *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998, doi: [10.1038/30918](https://doi.org/10.1038/30918).

[8] K. Faust and S. Wasserman, “Blockmodels: Interpretation and evaluation,” *Soc. Networks*, vol. 14, no. 1, pp. 5–61, Mar. 1992, doi: [10.1016/0378-8733(92)90013-W](https://doi.org/10.1016/0378-8733(92)90013-W).

[9] H. Tuomisto, “A consistent terminology for quantifying species diversity? Yes, it does exist,” *Oecologia*, vol. 164, no. 4, pp. 853–860, Dec. 2010, doi: [10.1007/s00442-010-1812-0](https://doi.org/10.1007/s00442-010-1812-0).

[10] J. Memmott, N. M. Waser, and M. V. Price, “Tolerance of pollination networks to species extinctions,” *Proc. Biol. Sci.*, vol. 271, no. 1557, pp. 2605–2611, Dec. 2004, doi: [10.1098/rspb.2004.2909](https://doi.org/10.1098/rspb.2004.2909).

[11] M. Park, E. Leahey, and R. Funk, “Dynamics of disruption in science and technology.” 2021, [Online]. Available: <https://arxiv.org/abs/2106.11184>.

[12] M. M. Dankulov, R. Melnik, and B. Tadić, “The dynamics of meaningful social interactions and the emergence of collective knowledge,” *Sci. Rep.*, vol. 5, p. 12197, Jul. 2015, doi: [10.1038/srep12197](https://doi.org/10.1038/srep12197).