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

Tuan Pham[1]

The University of Chicago, Chicago IL, USA
tuanhpham@pm.me

Abstract. This paper presents a computational model of acquiring new knowledge through self-learning (e.g. a Wikipedia "rabbit hole") or social influence (e.g. recommendations through friends). This is set up in a bipartite network between a static social network (*agents*) and a static knowledge network (*topics*). For simplicity, the learning process is singly parameterized by α as the probability of self-learning, leaving $1 - \alpha$ as the socially-influenced discovery probability. Numerical simulations show a tradeoff of α 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 α, or increased self-learning tendency, lead to higher population diversity and robustness. However, lower values of α, where learning/discovery is more easily influenced by social inputs, expand individual knowledge diversity more readily. These numerical results might provide some basic insights into how social influences can affect the diversity of human knowledge, especially in the age of information and social media.

Keywords: topic diversity, knowledge discovery, bipartite network, social influence, self-learning

1 Introduction

As the world becomes more connected and the amount, as well as accessibility, of information in it increases, how do we learn an existing body of knowledge, while simultaneously updating with the new, incoming information? How diverse is knowledge acquired through social interactions? Coupled with limited cognitive capacity, do people become more specialized or generalized as a result, especially since specialization has implications for creativity and research productivity [9]?

Answering these questions might be difficult at this point without assessing simple cases of learning within static networks. Previous work addressed topic diversity analysis at the population level without considering the evolution of new topic acquisition [12], or used dynamic processes and analyses within only the intralayer networks [8]. A recent study discusses the involvement of social interaction in innovation dynamics, but does not directly address the level of social influence nor analyze the resulting knowledge diversity in details [5]. 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 could learn something new. This project focuses on only two ways: (1) active *self-learning* by acquiring new knowledge through related topics (Fig. 1b); and (2) through *social influence* as suggested by one's social circle (Fig. 1c). An example of the former is following a "rabbit hole", starting from an already-known topic on Wikipedia or a reference in a journal article's bibliography section. On the other hand, examples of the second scenario include movie recommendations from friends or new research papers shared via social media. As humans tend to have limited capacity for learning, how would 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 (e.g. can all topics be learnt)?

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 process tends to improve topic diversity in the population manner, while recommendations through social influence would generally benefit individual diversity. Consideration of groups within the models have mixed effects at the individual level more so than the population level.

2 Methods

2.1 Model

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, respectively, of the agent graph G_a and topic graph G_t (Fig. 1a). The bipartite incidence matrix τ 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 an update process (described below). 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. Each parameter set (α, intralayer models, interlayer initialization) were simulated 5 times.

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 α probability, by learning about the related topics of things an agent already knows about (Fig. 1b). On the other hand, with probability β (for simplicity assumed to be $= 1 - \alpha$), an agent could acquire a

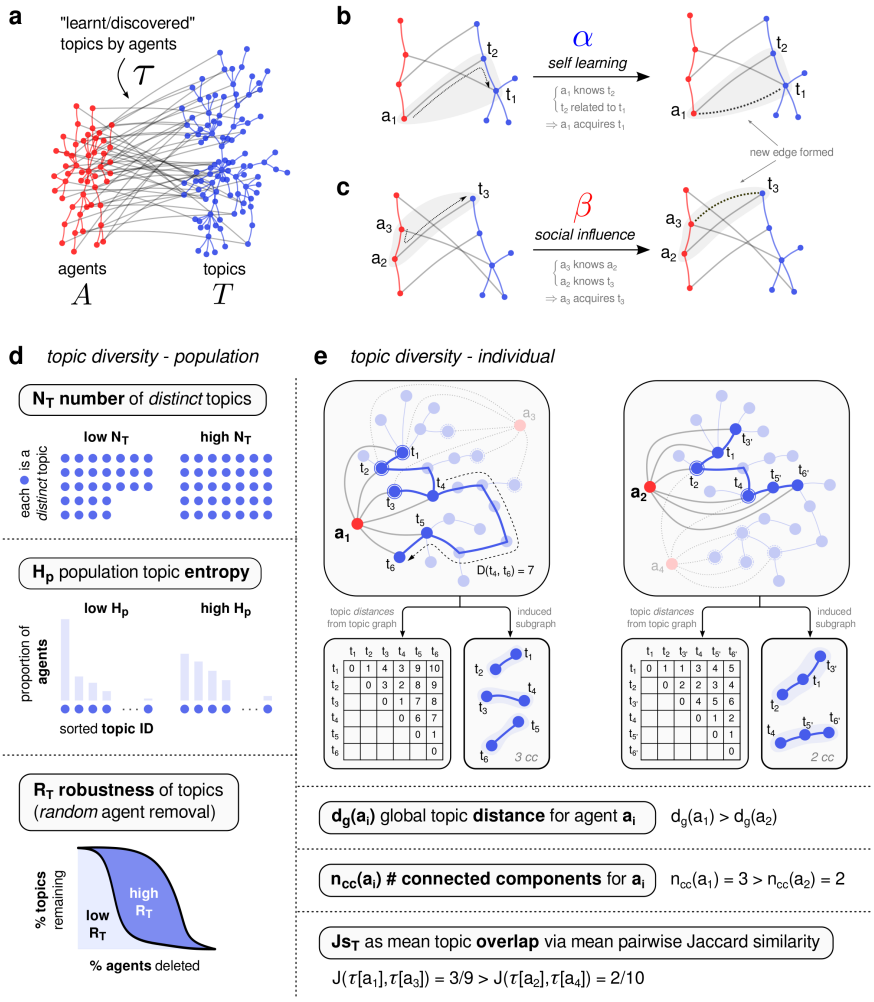

Fig. 1. *Description of the topic update/discovery process in the model and the different knowledge diversity metrics.* (**a**) Illustration of the intralayer agent graph (*red*) and topic graph (*blue*) with the interlayer edges (*gray*) representing the knowledge set of the agents. *Gray* triangles 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 for detailed descriptions).

new topic edge by traversing its neighbors in the agent graph then to the topic graph (Fig. 1c). One way to implement this is below.

Define $\psi(X)$ as a column L1 normalization operation on a matrix X, in which each column vector $\mathbf{x}_i$ of the matrix is normalized to $\mathbf{x}_i/||\mathbf{x}_i||_1$. 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 τ) with its column vector $\mathbf{p}_i$ defines the probability agent a_i chooses a new topic. A way to define this probability is:

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

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

The multiplication steps perform the traversal through neighbors across the intralayer networks. The binarization and subtraction with the current τ simplifies the implementation, so that agents only learn new topics and avoid "being stuck" around too popular topics. Additionally, for simplicity I consider $\beta = 1-\alpha$ so the process is only defined by α. 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 hyperparameters (except only for the number of nodes) are the same for agent and topic graphs for each simulation.

Nonblock models: The first approach is non-block networks. In the main text, only the scale-free (SF) network models are discussed, constructed by the linear preferential attachment models (`PA`) [1] (see Fig. 2). Additionally, other models are also analyzed: nonlinear `PA` models, Erdős–Rényi (`ER`) networks [3] with different connectivity probability, as well as small-world networks generated with the Watts–Strogatz (`WS`) models [11] (see Fig. S1a).

Block models: Since there are usually communities in real-world networks, (researchers or papers within the same field), the stochastic block models (`SBM`) [4] are used to emulate this phenomenon 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_{within}) or between groups (p_{between}). For simplicity, the former is kept fixed while varying the latter (see Fig. 3).

Interlayer initialization At the initialization stage, the probability of connection between a given agent and topic is uniform across topics. However, it is possible that other initialization strategies could influence the results. Hence, two different interlayer initialization strategies are introduced, 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 depend on its degree in G_t. A way to do this is to perform the $\texttt{softmax}\left(\{d_i\}; \beta_\sigma\right)$ on the degrees, basically transforming the degree sequence $\{d_i\}$ to a probability distribution. With $\beta_\sigma < 0$ ($\texttt{SOFTMAX}_1$), low degrees are favored; $\beta_\sigma = 0$ is equivalent to random initialization ($\texttt{SOFTMAX}_2$), while $\beta_\sigma > 0$ ($\texttt{SOFTMAX}_3$) favors high degree topics (Fig. S2a).

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_{sg} (Fig. 3) 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.

2.2 Diversity metric

These diversity metrics are illustrated in Fig. 1d,e.

Population Three population indices are defined, taken from an ecological perspective [10]. First, N_T is the number of distinct topics discovered when taking into account all agents' learnt topics, where higher values correspond to higher diversity. Second, H_p is the topic population entropy – the Shannon entropy from the discrete probability distribution of all the topics in the population (again, higher would suggest more diversity). Lastly, inspired by ecological network stability analysis [6], 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 R_T, higher values of which indicate that many agents need to be deleted 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, 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 $\tau[a_k] = \{t_h | \tau[t_h, a_k] = 1\}$ then $d_g(a_k) = \langle D(t_i, t_j; G_t) \rangle_{t_i, t_j \in \tau[a_k]}$. Higher values suggest that the agents know more outside of their comfort zone and tend more towards generalists. Another metric is the number of connected components $n_{cc}(a_k)$ in the induced subgraph $G_t(\tau[a_k])$, where higher values indicate there are many "islands" of topics that the agent knows about, again leaning towards generalist. Lastly, the mean pairwise Jaccard similarity Js_T between agents' topic sets are calculated, lower values suggest higher local diversity.

Group With groups in the block intralayer models, I also calculated the entropy of the topic group distribution, in both the population sense H_{gp} and individual sense H_{gi}. More specifically, H_{gp} is the entropy of the 10 topic groups considering the group identities of all topics learnt by all agents. On the other hand, H_{gi} is the average entropy of each agent's individual topic entropy. These two quantities are different; e.g. H_{gp} can be maximized (all groups uniformly distributed) while H_{gi} is 0 (each agent only learns about the topics within the same group).

2.3 Code availability

The source code is at `https://github.com/tuanpham96/topic-diversity`. The simulations were run in parallel on Azure VM. Simulations, analyses and visualizations were performed in R, further illustrated in Inkscape.

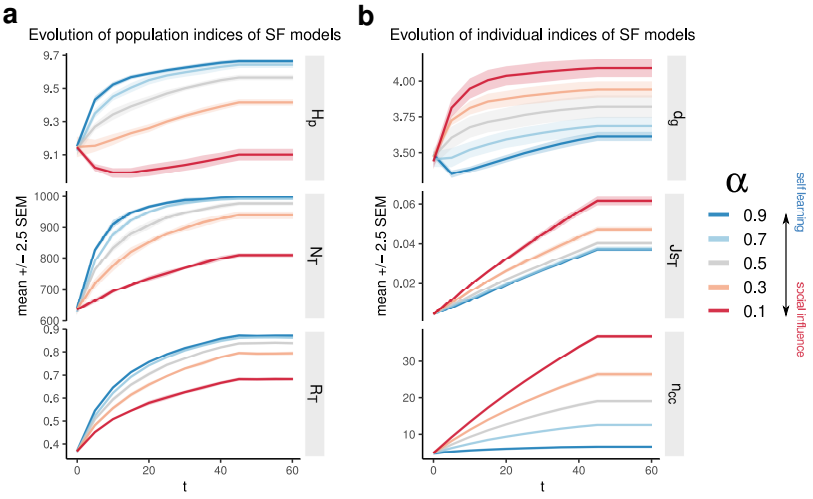

Fig. 2. *Changes in population topic diversity indices* (**a**) *and individual diversity indices* (**b**) *of the scale-free (SF) intralayer models due to* α (colors). H_p: topic population entropy; N_T: number of distinct topics; R_T: robustness to *random* removal of agents; d_g: mean distance of 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 Sect. 2.2 and Fig. 1d,e for more details. Each line plots the mean changes of 5 realizations for each index, analyzed every 5 steps.

3 Results

3.1 *Nonblock* intralayer models

The changes of the different diversity metrics for the scale-free networks are shown in Fig. 2 as an example to illustrate the tradeoff effect of self-learning versus social influence probability on population and individual diversity.

Generally, topic population diversity increases with self-learning probability α in terms of the topic entropy H_g and number of topics N_T. Through learning/discovery over time, low α could still achieve better population diversity. However, it does not seem likely for the worst case considered here, where entropy does not even increase considerably past 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 reasonable to say that higher α improves topic population diversity. Additionally, higher α leads to more robust retainment of the topics under random agent removal (i.e. higher R_T), though this is possibly partially an effect of higher N_T.

On the other hand, topic individual diversity usually decreases based on the chosen metrics. Increased α leads to decreased mean learnt topics distance d_g

and number of components n_{cc} in the induced subgraphs. Intuitively, higher social influence – lower α – would allow the agents to access topics outside of their comfort zone more easily, hence their own subgraph of topics tend to be more generalized, whereas higher α leads to more specialization. Lastly, at the local level Js_T, lower α 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 τ.

These trends are qualitatively consistent across different considerations of non-block models (Fig. S1), albeit some quantitative differences (especially in individual indices). Increases in α lead 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}). Under consideration of 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 generally detrimental across different population, local and individual diversity indices, especially for those 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. S2b).

In summary, in non-block intralayer models, higher self-learning (higher α) leads to higher topic diversity in a local and population context, but higher social influence (lower α) encourages individual topic diversity. Initializations that favor more popular topics have a negative effect on these different metrics.

3.2 *Block* intralayer models and topic group diversity

As real-world networks usually contain communities within them, the stochastic block intralayer models (`SBM`) are used to model intralayer networks and investigate how diversity indices change due to α and network modularity. Generally, the trends for population diversity and robustness during the simulation are similar to those previously discussed (Fig. S3a). Such trends as a function of model modularity do not differ much. Looking at the group population entropy H_{gp} (Fig. S3b, bottom), only when the networks are less modular do those values show a difference, albeit very small.

In the individual perspective (Fig. S3c), I note that group modularity increases diversity indices d_g and n_{cc}, possibly because there are fewer long-range links. The trends for local diversity are roughly similar and not affected much 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. S3b, 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 α do not seem to matter any more.

The final values of these different metrics in Fig. 3, with different group-correspondence initialization strategies, reveal these effects more clearly.

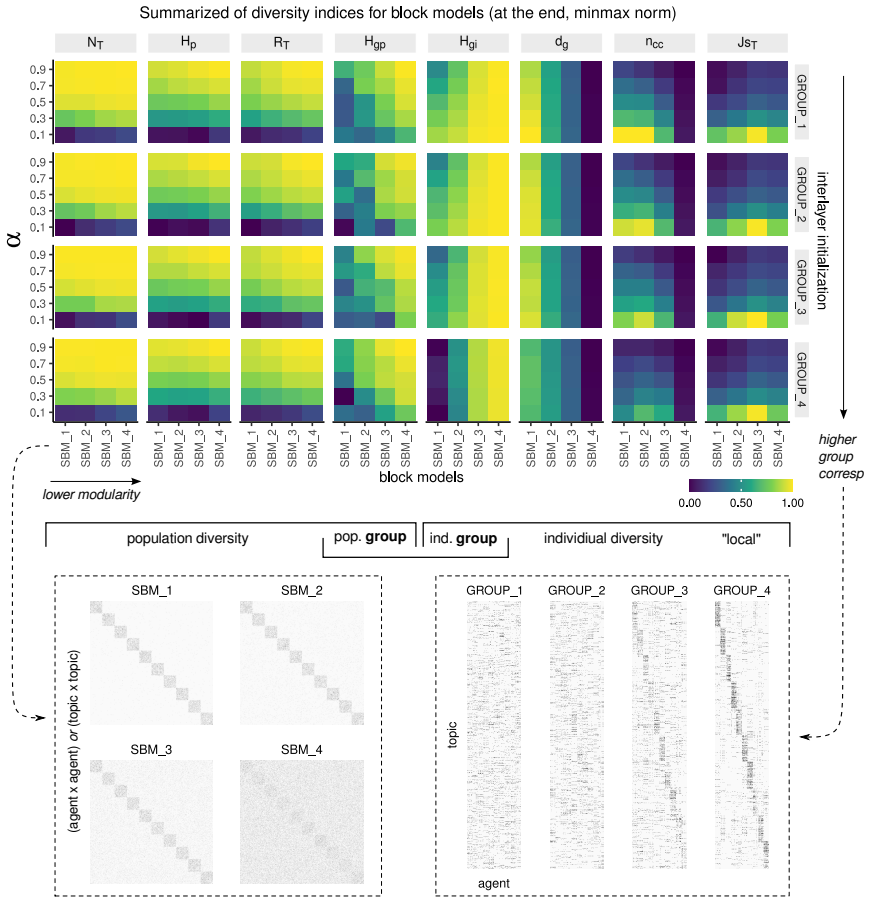

Fig. 3. *Summary of population and individual diversity indices due to α, across different block models.* Within each heatmap, the x-axis shows decreasing modularity of intralayer model (via increasing inter-modular connectivity), and the y-axis is α. The color represents values at the end of the simulations, 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.

More specifically, higher α and lower intralayer modularity generally leads to higher 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 individual diversity (d_g, n_{cc}) 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 Js_T visibly. However, generally higher α and higher model modularity tends to decrease topic similarity, hence increasing local diversity.

For group entropies, low group modularity generally increases both topic group population (H_{gp}) and individual (H_{gi}) diversity. High initial correspondence seems to benefit group *population* diversity (although such benefits may be small, see Fig. S3b), but tends to decrease group *individual* diversity.

In summary, with consideration of intralayer block models, higher α (self-learning tendency) benefits more for population diversity and robustness, but less so for 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 α 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 could be harmful for these metrics. At the local level, higher α and high network modularity generally decrease knowledge 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 in bipartite networks, several interesting results emerge. First, increasing α (self-learning, traversing through interlayer edges first) leads to higher topic population diversity and robustness in several random models for the intralayer networks, including block and non-block models. However, such increase has drawbacks for topic individual diversity, as it reduces the chance for the agent nodes to acquire interlayer edges from topics that are usually distant from their comfort zone. Social influence, traversing through intralayer edges first (β route), would better influence individual diversity.

When intralayer groups are considered, group modularity may hurt the population diversity (some more than other) and more apparently the group individual entropy. Interestingly, it appears more beneficial for individual indices in term of graph distances and components. Although initial group correspondence does not have much effect on the population diversity, it has a dramatic drawback at the individual level (both entropy and graph metrics).

Limitations and future considerations Though there are interesting results in a theoretical sense as a toy model, there are many limitations with the current study. One of the major drawbacks is the assumption that the knowledge space would eventually outgrow the agent population size hence considering only $n_t > n_a$. Comparing the current estimate size of the indexed WWW [1] to the total internet users [2] might satisfy this assumption. However, a more realistic approach would be to consider the subset of the internet involved directly with knowledge, e.g. Wikipedia. Then the size of the topic network [3] is around 2-3 orders of magnitude *smaller* than the the potential agent network [2], countering

[1] 30-50 billion as of Oct 3, 2021 (`worldwidewebsize.com`)

[2] 4.66 billion internet users (`statista.com/statistics/617136/`)

[3] 54.3 million total pages (6.4 million English) (`wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia`)

this assumption. Another assumption is that the intralayer models for the topic and agent graph are considered similar for simplicity and for practical reasons.

Future studies need to address these issues by integrating realistic knowledge and social network networks (or sampled versions of them) into the model, or at least the intralayer size ratios and generative models based on realistic network statistical properties. Other considerations include: different ratios between intralayer network sizes; inclusion of directed weighted edges (strengths could imply confidence in knowledge in τ); non-persistent interlayer edges; different update probabilities (serendipity, forgetting, mastering, ...); the cost of learning new subjects; delays in acquiring new knowledge; different versions of the update equation; the decreased disruptiveness in new knowledge discovery [7]; and, more importantly, the dynamic nature of the intralayer networks (e.g. [8]). Furthermore, future endeavours should also 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 (e.g. Twitter followers and hashtags) [12].

Additionally, further analyses should include examination of: the modularity changes in the bipartite τ [2] or in the projected graphs (for example, low α 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); relationships between an agent's properties (e.g. degree, centrality) & position (e.g hub, periphery) in G_a and the properties of their acquired topic sets (e.g. whether hub agents are more likely to also obtain peripheral arcane knowledge than peripheral agents); and lastly, persistent homology analyses (since the interlayer edges are defined as persistent here).

5 Acknowledgement

I would like to thank the reviewers of CNA 2021 for their comments, Dr. Mercedes Pascual and Dr. Sergio A. Alcala Corona for their *Networks in Ecology and Evolution* course (University of Chicago), Dr. Julie S Haas (Lehigh University) and my friends Sam Nguyen, Poojya Ravishankar, Silas Busch for their discussion and feedback on the model, interpretations and writing. I would also like to acknowledge the Graduate Council Research & Personal Development Fund (University of Chicago).

References

1. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509–512 (Oct 1999)
2. Dankulov, M.M., Melnik, R., Tadić, B.: The dynamics of meaningful social interactions and the emergence of collective knowledge. Sci. Rep. 5, 12197 (Jul 2015)
3. Erdös, P., Rényi, A.: On random graphs I. Publicationes Mathematicae 6, 290–297 (1959)

4. Faust, K., Wasserman, S.: Blockmodels: Interpretation and evaluation. Soc. Networks 14(1), 5–61 (Mar 1992)
5. Iacopini, I., Di Bona, G., Ubaldi, E., Loreto, V., Latora, V.: Interacting discovery processes on complex networks. Phys. Rev. Lett. 125(24), 248301 (Dec 2020)
6. Memmott, J., Waser, N.M., Price, M.V.: Tolerance of pollination networks to species extinctions. Proc. Biol. Sci. 271(1557), 2605–2611 (Dec 2004)
7. Park, M., Leahey, E., Funk, R.: Dynamics of disruption in science and technology (2021)
8. Sun, Y., Latora, V.: The evolution of knowledge within and across fields in modern physics. Sci. Rep. 10(1), 12097 (Jul 2020)
9. Teodoridis, F., Bikard, M., Vakili, K.: Creativity at the knowledge frontier: The impact of specialization in fast- and slow-paced domains. Adm. Sci. Q. 64(4), 894–927 (Dec 2019)
10. Tuomisto, H.: A consistent terminology for quantifying species diversity? yes, it does exist. Oecologia 164(4), 853–860 (Dec 2010)
11. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393(6684), 440–442 (Jun 1998)
12. Weng, L., Menczer, F.: Topicality and impact in social media: diverse messages, focused messengers. PLoS One 10(2), e0118410 (Feb 2015)

Supplementary figures

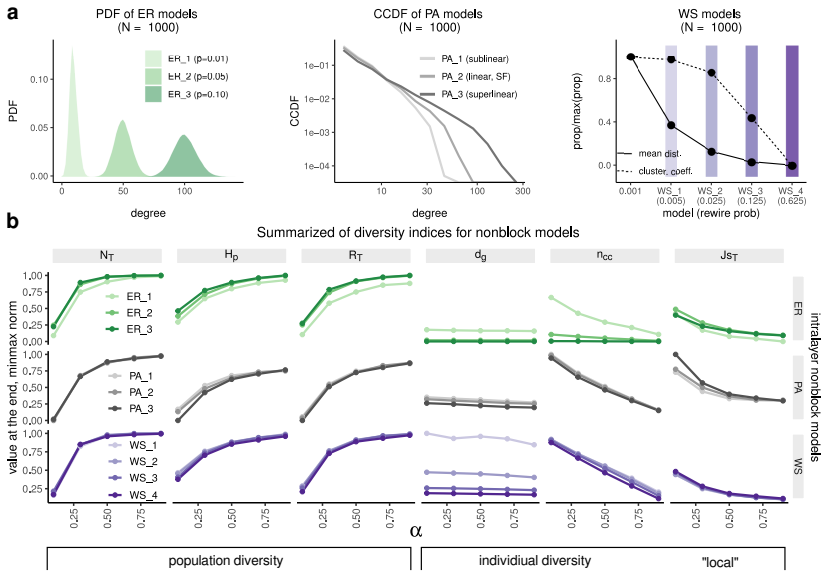

Fig. S1. *Variations of nonblock intralayer models.* (**a**) Set up of nonblock models. `PA`: preferential attachment, `ER`: Erdős–Rényi, `WS`: Watts–Strogatz (Sect. 2.1) (**b**) Changes of diversity indices for these models as a function α (Sect. 2.2)

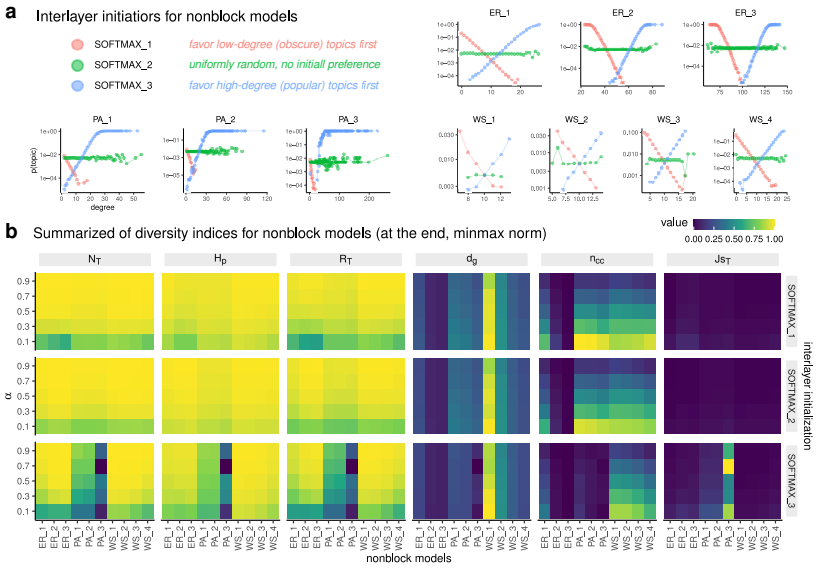

Fig. S2. *Different initialization strategies for nonblock models* (**a**) *based on the topic intralayer degrees and* (**b**) *effects on population and individual diversity indices as a function of* α. See Fig. S1a for names and illustrations of the different models.

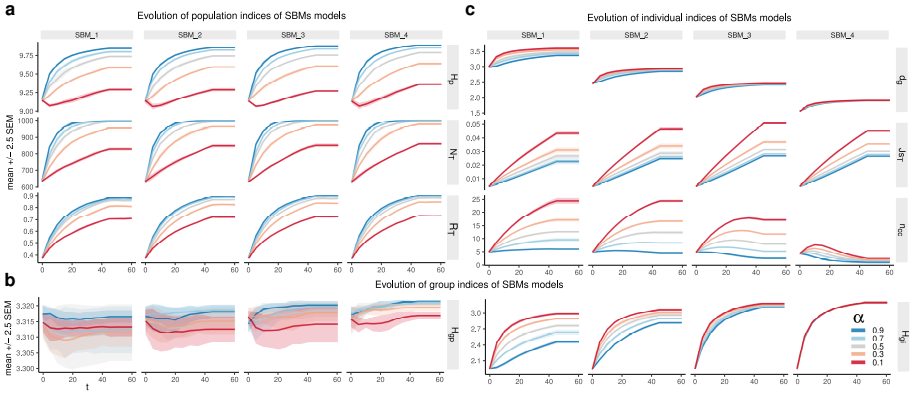

Fig. S3. *Changes of population diversity indices* (**a**) *, group diversity indices* (**b**) *and individual diversity indices* (**c**) *for the stochastic block intralayer models due to* α.