From Overload to Insight: A Network Science Approach to Personalized Literature Review

Timothy James

Department of Applied Mathematics, Naval Postgraduate School,
1 University Cir. Monterey, CA 93943, USA;
timothy.james@nps.edu,

Abstract. How can we personalize the discovery of research to support scholarly reading? The rapid expansion of scholarly literature across disciplines presents a growing challenge for researchers, educators, and students striving to stay informed and to make meaningful contributions. This work introduces a network-based methodology for identifying what papers in a research area a researcher should read. By leveraging topological metrics such as k-core and community detection, we offer a scalable and objective framework for identifying the communities that form the body of work and their representative papers. Building on this foundation, we propose a personalized literature review methodology that, based on a user's academic background and chosen keywords, recommends the most relevant papers to read.

Keywords: network science, personalized literature review, k-core

1 Introduction and Motivation

The volume of academic literature is growing at an unprecedented pace. Challenges to remain current and time constraints make it increasingly difficult for researchers to identify which publications are most relevant to their research questions. Citation count and impact factor metrics provide global perspective, but they may overlook foundational or newer state-of-the-art publications within specific subfields. As a result, even experienced researchers may miss key studies, and newcomers may struggle to orient themselves within a field [1, 5, 7]. To address this challenge, we introduce a network-based approach to identify and summarize literature review based on keyword searches. We call the citation network induced by our keyword search the Keyword-Induced Literature Network, KN. By analyzing the KN's topology using metrics such as k-core and community detection, we move beyond high-level indicators to uncover the hidden architecture of scholarly influence in the researcher's area of interest. Furthermore, we propose ideas for a personalized literature review methodology that tailors reading recommendations to individual users based on their academic background and keyword queries. This methodology enables researchers to focus their attention on the studies most relevant and impactful to them, thereby enhancing the efficiency of literature reviews and supporting more informed, strategic engagement with the scholarly landscape.

2 Existing work

Similar approaches have been used to examine KNs obtained from PubMed, a free website for scholarly research restricted to medical articles [6]. Figure 1 shows an example of the k-core degeneracy from an analysis by Dinkel of such a KN [4]. The x-axis shows the value k of the k-core analyzed, and the y-axes shows the count of nodes (papers), edges (citations), and communities (clusters of papers in KN) in that k-core.

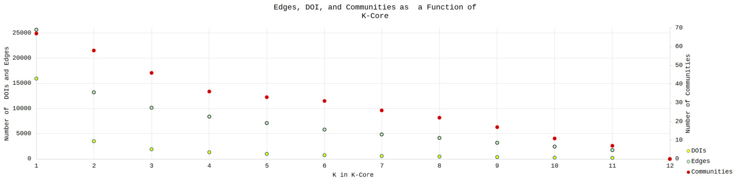

Fig. 1. Count of edges, nodes, and communities as a function of k-core [4]

As expected, Figure 1 shows that the counts drop as k increases. Dinkel's work is based on analyzing the core of this KN, namely the 11-core [4]. We seek to understand if there is an optimal stopping criteria, that balances reducing the number of papers identified for a literature review and preserving community representation. i.e., does it make more sense to stop at a lower k value?

Connecting learning science and cognitive theory, *Flow* theory is a state of deep immersion and enjoyment where one is fully engaged and focused. It is characterized by a balance between the challenge of the task and the cognitive ability of the individual. Dynamic, adaptive systems can be helpful to guide an individual back into the *Flow* when they feel frustrated or bored [2].

3 Proposed Solution: The Synthesized Network

To reduce the overall computational cost of potential personalization algorithms, we seek to identify the optimal k value for each KN. Consider the forward difference, $\Delta f(k) = f(k+1) - f(k)$, the analog of the derivative for discrete functions, of the communities vs k-core function. We identify local minima in the forward difference and select the k-core associated with the first local minimum of $\Delta f(k)$ where the number of communities represented, N_k, is less than the maximum number of papers the researcher is willing to read, H. Figure 3 shows the forward difference for the plot of communities vs k-core from Figure 1.

We introduce the Personal Profile, PP, by capturing and vectorizing a researcher's educational background, the number of papers they are willing to read (a range between L and H), keywords of interest, and keywords to avoid. Concurrently, we introduce the concept of the Synthesized Network, SN, as the k-core we select on which to conduct the synthesis of the literature review.

We propose an algorithm to compute the distance from the PP to the vectorization of each paper in the SN. We vectorize AI-generated summaries of each abstract in the SN and calculate the semantic similarity between them

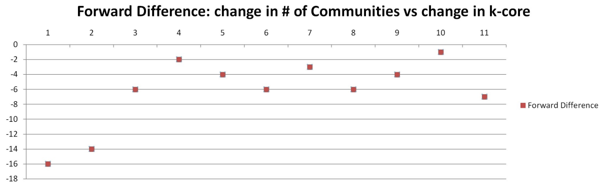

Fig. 2. Forward difference of communities in Figure 1

and the PP [3]. This represents the researcher's ability with respect to each paper, assigning a difficulty rating based on the type of paper and where it was published. Let N_{sn} be the number of communities remaining in the SN; we propose a process that selects the p_i papers in each community c_i for which the difficulty most closely matches the researcher's ability. We identify those papers in $1 \leq i \leq N_{sn}$ as most closely matched in SN to the researcher's PP. We propose a weighted distribution of papers based on the order $n_i \propto |V(c_i)|$ of remaining communities in SN. The number of papers recommended to read in each remaining community is $p_i = L * P_i = \frac{Ln_i}{\sum_{j=1}^{N_{cs}} n_j}$

Finally, we propose data collection efforts focused on areas where subject matter experts can provide reliable validation of our results.

4 Preliminary Experimental Evaluation

Suppose there is a researcher who indicates in their PP that they are willing to read between $L = 20$ and $H = 50$ papers for a literature review. Observe that the k-core associated with the first forward difference local minimum where $N_k < H$ is the 6-core. Hence we identify the 6-core as the SN, having 748 of the original 15,977 papers (less than 5%) but still $N_6 = 27 < H = 50$ of the original 69 communities (nearly 40%) represented. The 6-core SN is shown in Figure 3.

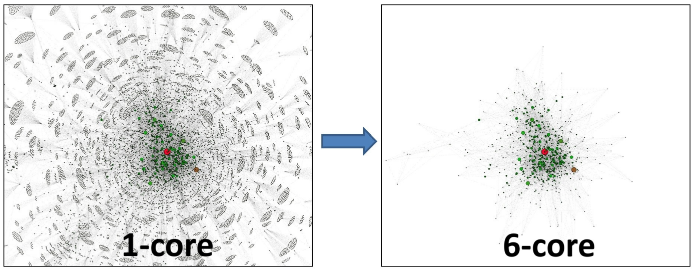

Fig. 3. Visualization of the network (1-core) & its Synthesized Network (6-core)

Table 1 shows the resulting number of recommended papers to read for this experiment. Columns not shown contain a single recommended paper in the respective community.

We rank order, by difficulty, the p_i papers for which difficulty and ability are most closely matched (minimal distance from $y = x$ in the *Flow* plot) (see Figure 4).

Table 1. Recommended Number of Papers to Read in each Community

Community Label (c_i)	25	29	9	6	35	46	...
Number of Papers in Community (n_i)	225	174	158	66	16	15	...
Percent of Papers in Community (P_i)	30.1%	23.2%	21.1%	8.8%	2.1%	2.0%	...
Recommended Number of Papers to Read (p_i)	7	5	5	2	1	1	...

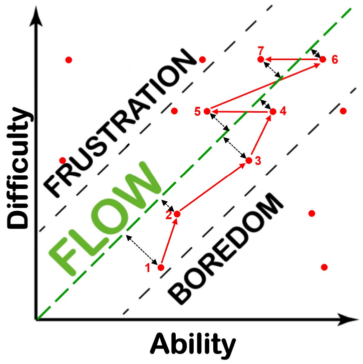

Fig. 4. Example of the *Flow* of research paper recommendations for Community 25

In this manner, we rank order the papers in each community and provide personalize recommended path through the literature review which meet the researchers desired volume of reading while maximizing community representation.

References

1. Bornmann, L., Haunschild, R., Mutz, R.: Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. Humanities and Social Sciences Communications **8**(1), 1–15 (2021)
2. Csikszentmihalyi, M., Abuhamdeh, S., Nakamura, J.: Flow, pp. 227–238. Springer Netherlands, Dordrecht (2014)
3. Diaz, D.O., Gera, R., Keeley, P.C., Miller, M.T., Leondaridis-Mena, N.: A recommender model for the personalized adaptive chunk learning system (2019)
4. Dinkel, B.E.: Beyond Citation Counts: Network-Based Strategies for Detecting Influential Medical Research. Master's thesis, Naval Postgraduate School, Monterey, CA (June 2025)
5. Foundation, N.S.: Publications output: U.s. trends and international comparisons. Tech. rep., National Science Foundation (2024), https://ncses.nsf.gov/pubs/nsb20214
6. National Center for Biotechnology Information: PubMed overview (2025), https://pubmed.ncbi.nlm.nih.gov/about/, accessed: March 1, 2025
7. Snyder, H.: Literature review as a research methodology: An overview and guidelines. Journal of business research **104**, 333–339 (2019)