Impact-Oriented Contextual Scholar Profiling using Self-Citation Graphs

Yuankai Luo, Lei Shi*, Mufan Xu, Yuwen Ji, Fengli Xiao, Chunming Hu, Zhiguang Shan

SIGKDD 2023

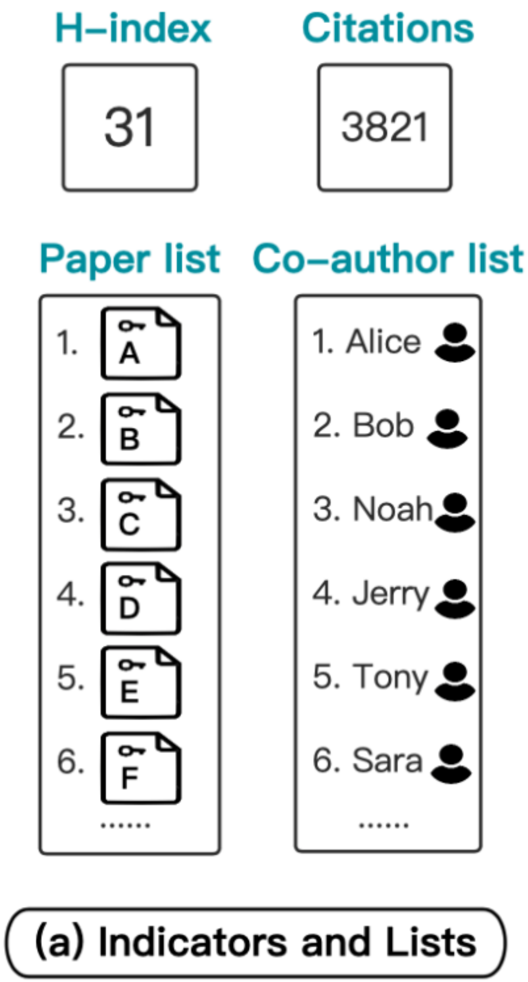
罗元凯
北京航空航天大学 计算机学院

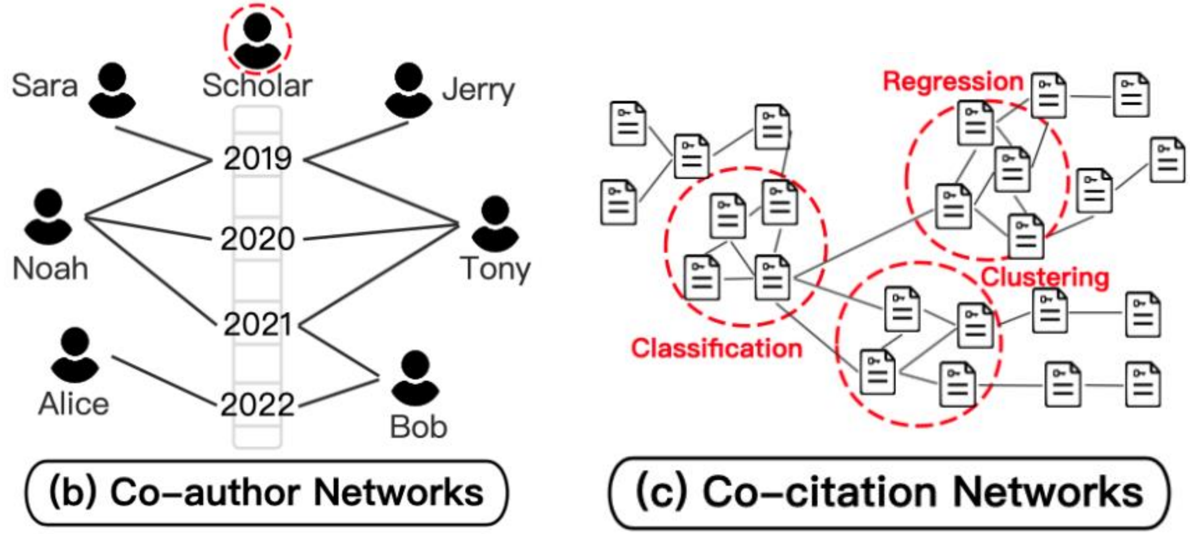
Outline

- Background
- GeneticFlow Framework
- Contextual Scholar Profiling
- Evaluation

Background

How to arrange a scholar's academic data to best represent his/her scientific impact?





Existing scholar profiling:

- (a) The standard indicator and list view exemplified by Google Scholar
- (b,c) Bibliometric networks:
- the co-authorship network
- the co-citation network

Background

Scholar profiling should consider the following requirements:

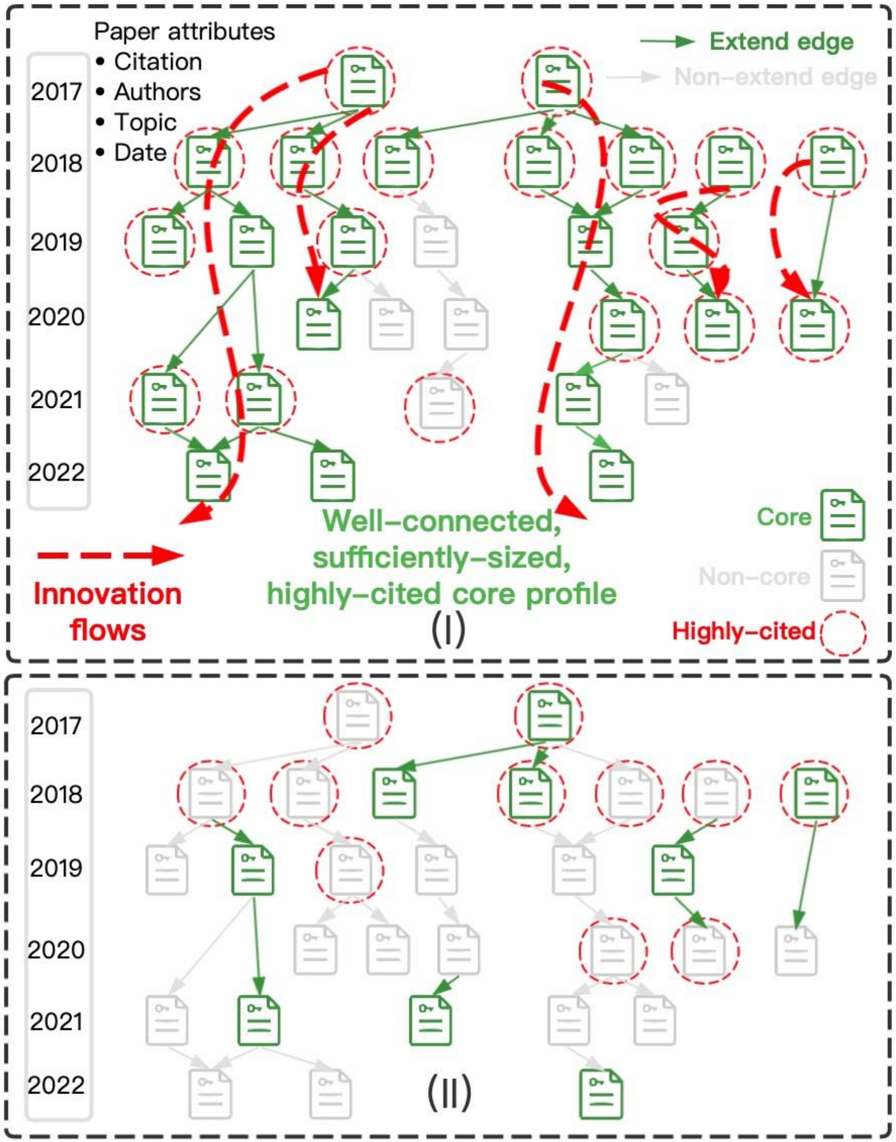
- (a) **Structured-context**: the complex academic data of a single scholar should be integrated into a structured representation.
- (b) **Scholar-centric**: the profile should focus on the target scholar only.
- (c) **Evolution-rich**: the profile should track the evolution of a scholar’s scientific impact.

Our idea: GeneticFlow

self-citation graph

effective in profiling the innovation flows of a scholar

GeneticFlow Framework



GeneticFlow (GF):

A timed, self-citation graph composed of all the papers with impact-oriented paper attributes. (structured-context profiling)

- Core paper: infer the set of core papers to be most representative to scientific impact.

- Core citation: detect the set of self-citations that truly represent the evolution.

Case:

I and II have the same citations, h-index, and paper count.

The scholar on the top is analytically of higher impact than the one on the bottom, with a well-connected, sufficiently-sized, and highly-cited core GF profile in the foreground.

Problem:

The problem is defined as finding the subgraph $G*$ of G that best represents the impact of the scholar.

How to find the subgraph?

Detect core paper s

The scholar should make a significant contribution to these papers.

- Assumption 1 (author order): A paper's contribution is unequally credited to all authors by author order unless the paper is alphabetically ordered.
- Assumption 2 (advisor-advisee credit sharing): An author's contribution to the paper is also credited to his/her advisor if only: a) the advisor is a co-author of the paper; and b) the advisor-advisee relationship is active at the publication date of the paper.
- Theorem (author contribution): On any paper v published at time t, the probability for the kth author a_k to contribute significantly can be estimated by

$$p_{cont}(a_k) = max(\frac{1}{k}, \max_{\forall l \neq k} \frac{p_{AA}(a_k, a_l, t)}{l})$$

- **Advisor-advisee detection:**

The advisor of an advisee in a research field at time t is characterized as an experienced researcher in the field (D1),

who supervised a sufficient number and ratio of major papers by the advisee (D2)

in a sufficiently long time (D3)

on the early career of the advisee in the field (D4).

$$p_{adr}(a_k, a_l, t) = \frac{N_{a_k}(0,t) - N_{a_k,a_l}(0,t)}{N_{a_k,a_l}(0,t)}$$

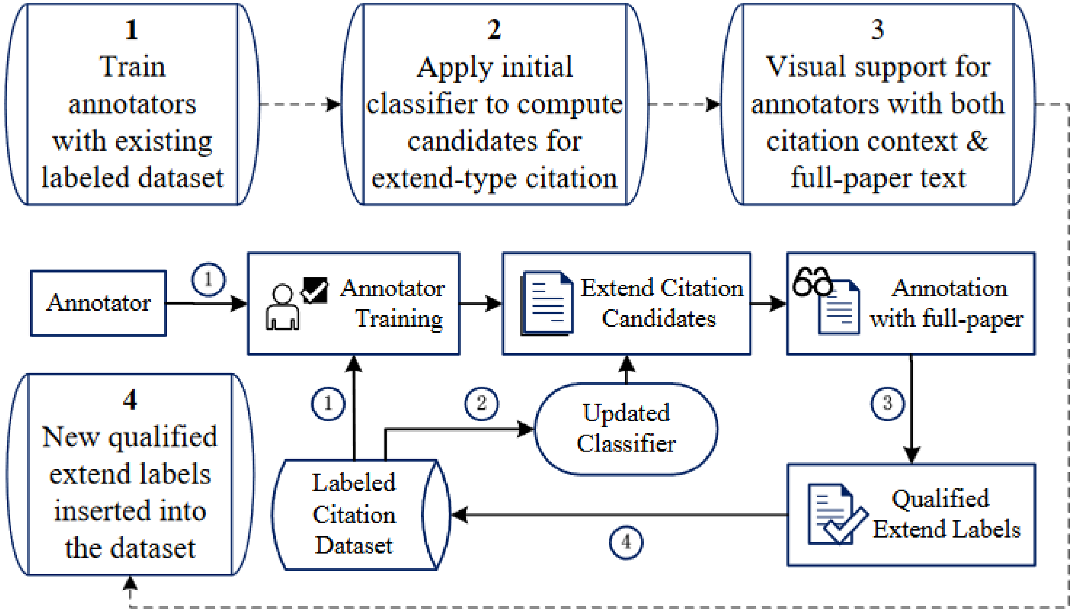
$$p_{ade}(a_k, a_l, t) = \max_{\substack{t_0 \le t \le t_1,\ t_1 - t_0 \ge S_{len} \\ numerator \ge S_{adr}}} \frac{\sum_{t_0 \le t \le t_1} \hat{N}_{a_k,a_l}(t) Mod(t)}{\hat{N}_{a_l}(t_0, t_1)}$$

$$p_{AA}(a_k, a_l, t) = min(1.0, p_{adr}(a_k, a_l, t)) \times min(1.0, p_{ade}(a_k, a_l, t))$$

Detect core citations (extend-type citations)

The author uses cited work as basis or starting point. And the new work will probably be an evolution of the scholar's research ideas.

- We use the supervised learning to infer core citations. We manually annotate extend type citations and create the training dataset.



Standard annotation process

leading to a dataset of 222/1604 positive/negative extend-type citation samples.

Contextual Scholar Profiling

- Hand-craft four categories of 20 features (Ft) interpretable for extend-type citation inference.

Category	Name	#Ft	Description	Sig.	Dataset
Paper-meta	# of citations_cited	1	citation count of the cited paper	0.0016	MAG
	year_diff	1	publication year difference between cited and citing papers	0.00027	ARC & MAG
	# of shared_authors	1	the number of shared authors between cited and citing papers	1.9e-48	ARC & MAG
Cite-net	co-citation	2	co-citation metrics between cited and citing papers	≤1.0e-07	ARC & MAG
	bib-coupling	1	bibliographic coupling metrics between cited and citing papers	5.2e-08	
Temporal	cross-correlation	3	cross-correlations between citation time series of cited/citing papers	≤0.037	MAG
Content	content-similarity	1	cosine similarity between vectorized content of cited and citing papers	1.3e-16	ARC
	# of cite_occurrences	1	the number of total occurrences of in-text citations of this citation link	4.0e-09	
	# of cites_occur_sec	3	# of cite_occurrences in key sections	≤0.044	
	cite_relative_pos	4	position of in-text citations in paper, section, sub-sec., sentence	≤0.049	
	lexical_pattern	2	appearance of certain phrases: "an/the extension", "our previous", etc.	≤5.0e-11	

- Performance of extend-type citation inference using various classifiers, feature sets, and the comparison with literature. We select the Extra-Tree model as the final classifier.

Metric	Classifier			Ablation study				Previous result	
	Extra-trees	MLP	DNN	(-) Paper-meta	(-) Cite-net	(-) Temporal	(-) Content	[49][51][35] merged	Report in [35]
F1 score	**.646±.014**	.543±.018	.544±.015	.636±.007	.639±.010	.639±.005	**.471±.009**	**.418±.019**	.403±.029
AUC	**.902±.005**	.806±.016	.785±.014	.871±.009	.898±.006	.899±.005	**.796±.008**	**.841±.009**	.775±.017
ACC	**.924±.002**	.901±.004	.899±.004	.921±.002	.922±.001	.924±.002	**.895±.002**	.949±.001	**.976±.001**

Evaluation

- The proposed GF profiling method is mainly applied to MAG, which covers 237M papers from all science areas, 240M authors, and 1.63B citations.
- To validate the effectiveness of GF profiling, we consider the task of inferring major scientific award recipients.

CS sub-field	Awards (except ACM fellow & Turing award)	# of awardees (top-500 scholars)	Sample list	Full GF profile: # of nodes, edges		Core profile	
				Awarded (50)	Others (150)	Nodes	Edges
NLP-ARC	ACL Lifetime Achievement Award / Fellow	77	#1~#207	121±56,205±173	93±50,153±134	66.5%	12.8%
Database	SIGMOD Innovations Award	114	#1~#247	118±61,166±126	74±36,112±79	64.0%	12.8%
Security	SIGSAC Outstanding Innovation Award	81	#1~#208	138±79,190±167	123±66,145±105	65.5%	18.3%
DM	SIGKDD/ICDM Innovations/Research Award	108	#1~#235	169±136,305±392	133±66,233±181	65.9%	25.1%
HCI	SIGCHI Lifetime Research Award / Academy	117	#1~#251	113±61,160±145	99±51,135±94	63.8%	29.9%
SE	SIGSOFT Outstanding Research Award	56	#1~#369	81±41,86±85	69±35,67±52	63.6%	12.5%
TCS	SIGACT Donald E. Knuth Prize	127	#1~#239	114±47,215±170	99±43,202±150	N/A	N/A
PL	SIGPLAN PL Achievement Award	135	#1~#244	90±34,165±134	87±37,187±180	N/A	N/A

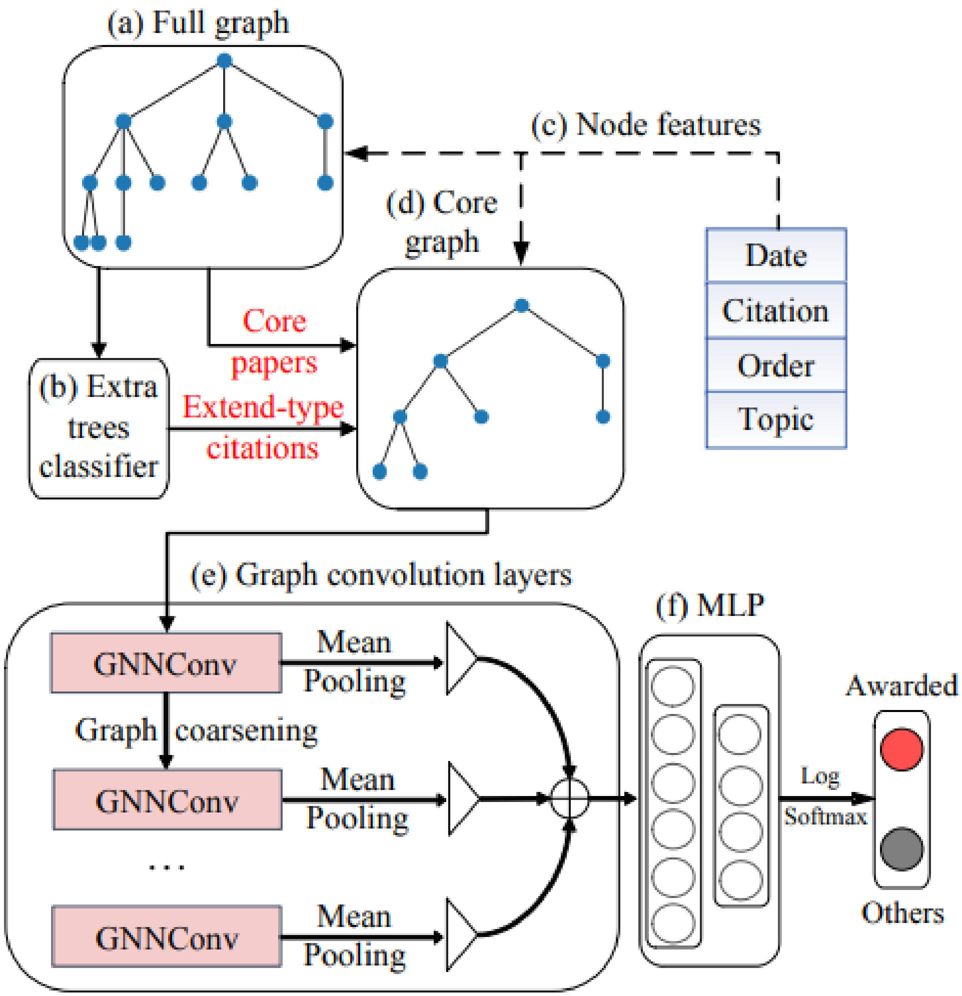
- We select 8 sub-fields of CS. In each field, we only consider the highest-class technical achievement/innovation awards plus ACM fellow and Turing award. And we sample 200 scholars including 50 award recipients and 150 other scholars.

Evaluation

- To apply GF methods to downstream tasks, we introduce graph neural network (GNN) models to learn high performance representation of profiling results.
- Node attributes:

the paper’s total citation count, the publication date,

the scholar’s order in the paper, the paper’s topic vector.

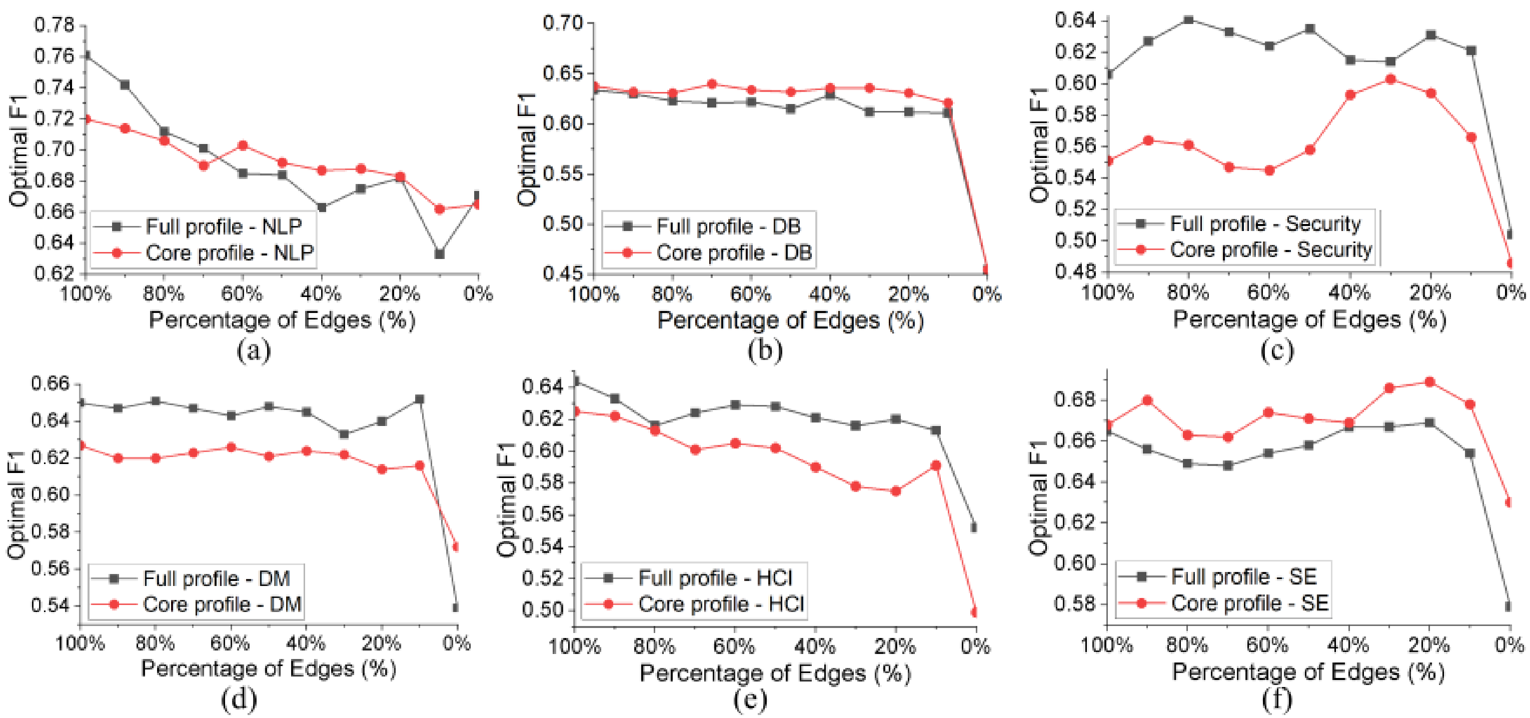


Evaluation

F1 measure in the award inference task using GeneticFlow and alternative methods.

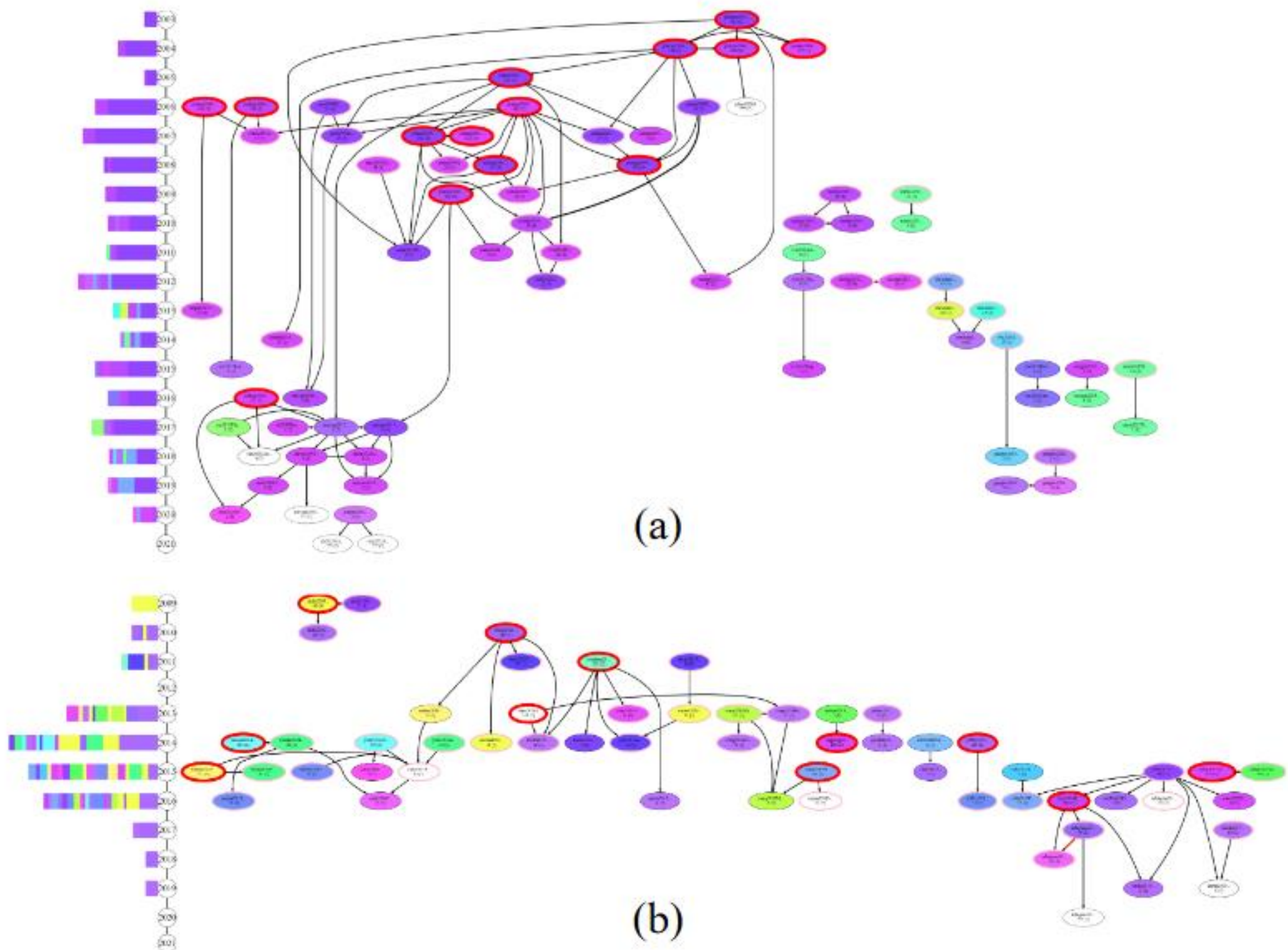
CS sub-field	GeneticFlow: Full profile	GeneticFlow: Best core profile	Author-Level Impact Indicators: SVM	Author-Level Impact Indicators: XGB	Author-Level Impact Indicators: RF	Author-Level Impact Indicators: MLP	Bibliometric Networks: CC	Bibliometric Networks: BC	Bibliometric Networks: CA
NLP-ARC	**.762±.016 (p<1e-4)**	.720±.018 (p=2e-4)	.632±.012	**.636±.013**	.621±.019	.629±.016	.531±.030	**.578±.021**	.473±.034
Database	.634±.018 (p=0.034)	**.638±.016 (p=0.012)**	.517±.020	**.546±.021**	.526±.020	.517±.016	.550±.021	**.588±.012**	.501±.035
Security	**.606±.020 (p=0.044[1])**	.551±.022	.557±.025	.572±.016	.548±.018	**.589±.021**	**.576±.017**	.572±.018	.528±.021
DM	**.653±.020 (p=0.007)**	.627±.014 (p= 0.045)	**.590±.012**	.533±.018	.574±.018	.574±.016	.563±.022	**.569±.019**	.476±.020
HCI	**.644±.018 (p=1e-4)**	.625±.016 (p=0.001)	**.562±.011**	.558±.017	.548±.025	.528±.017	**.551±.024**	.527±.022	.466±.029
SE	.665±.011 (p=0.023)	**.668±.009 (p=0.014)**	**.596±.011**	.558±.016	.512±.020	.593±.014	**.607±.023**	.595±.019	.523±.028

The performance of full/core GF profiles with varying edge percentages.



Evaluation

Case studies: https://vimeo.com/795348791/



(a)

(b)

Thanks for listening!