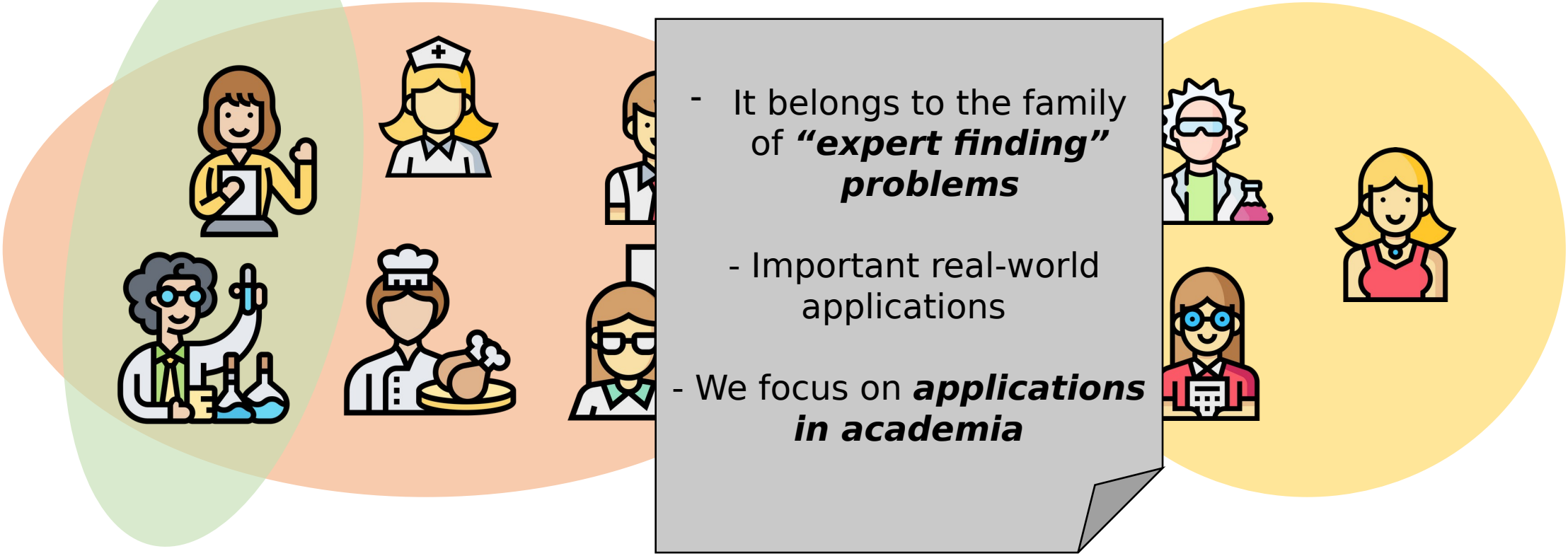
VeTo: Expert Set Expansion in Academia

Thanasis Vergoulis, Serafeim Chatzopoulos, Theodore Dalamagas and Christos Tryfonopoulos

Expert Set Expansion: The problem

C = candidate researchers

E_{kn} = set of known experts



- It belongs to the family of ***"expert finding" problems***

- Important real-world applications

- We focus on ***applications in academia***

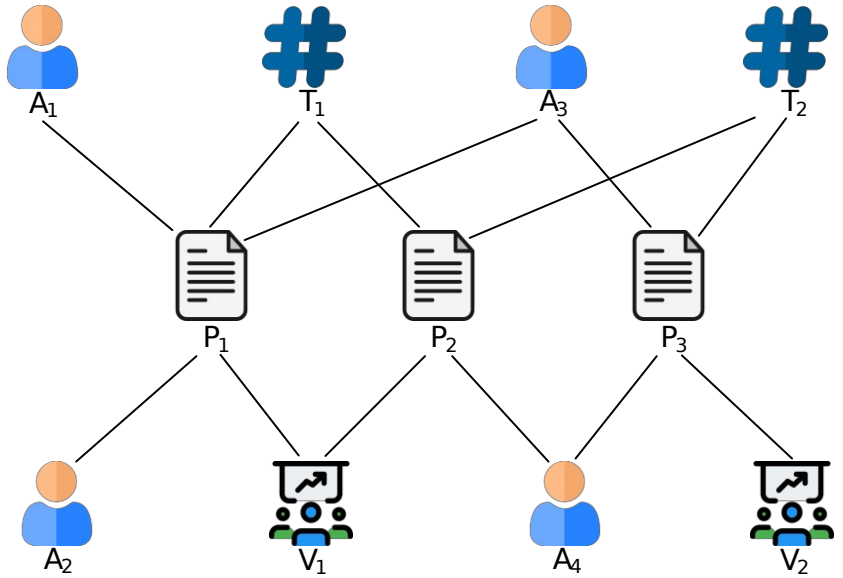
Find the n researchers in C that are most similar to experts in E_{kn}

Popular Expert Finding approaches

- **Topic Keywords approach**
  - Match topic to experts by utilizing ***topic keywords & person names in text corpora*** (e.g., publications, web pages).
  - Issues:
    - In many cases, ***difficult to describe topics*** as concrete sets of keywords
    - It relies on the availability of concrete text corpora – in academia full texts are often behind ***paywalls***
- **Querying by example approach**
  - Build ***expertise profiles*** (usually) based on analysing text corpora & ***match individuals based on these profiles***
  - Issues:
    - It relies on the availability of text corpora

Our approach

- Utilise ***Sholarly Knowledge Graphs***
  - ***Heterogeneous networks*** containing information about academia
  - Examples:
    - AMiner's DBLP-based datasets[1]
    - Open Research Knowledge Graph[2]
    - OpenAIRE Research Knowledge Graph[3]
- Very ***rich & relatively clean*** data
- Variety of data mining approaches to capture complex semantics
  - ***Metapath***-based analysis

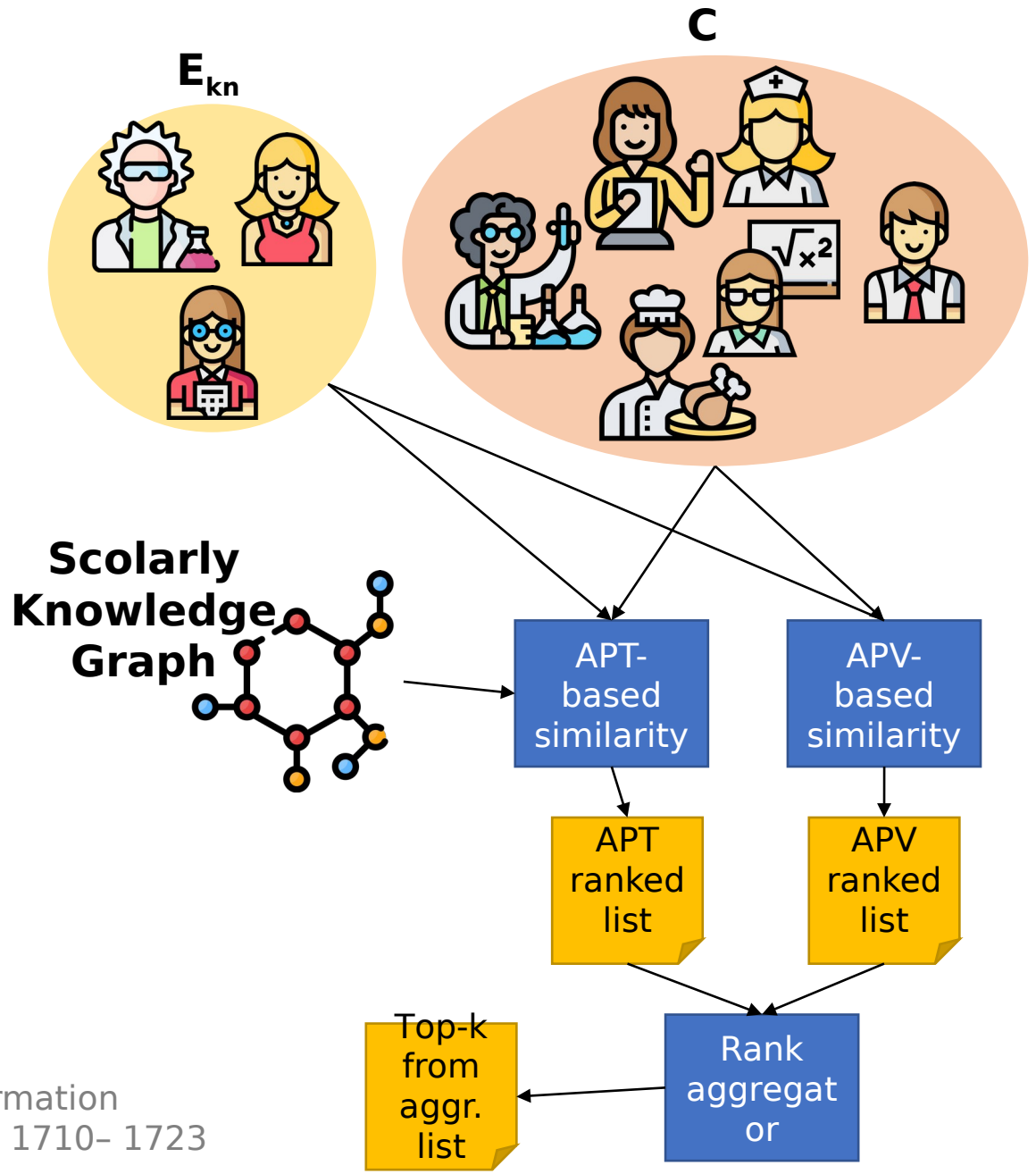


[1] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: Extraction and Mining of Academic Social Networks. In: Proceedings of the 14th ACM SIGKDD. pp. 990–998. ACM (2008)
[2] Jaradeh, M.Y., Oelen, A., Farfar, K.E., Prinz, M., D'Souza, J., Kismihó k, G., Stocker, M., Auer, S.: Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019. pp. 243–246 (2019). https://doi.org/10.1145/3360901.3364435
[3] Manghi, P., Bardi, A., Atzori, C., Baglioni, M., Manola, N., Schirrwagen, J., Principe, P.: The openaire research graph data model (Apr 2019).

VeTo

- It utilises ***metapath-based similarity of academics*** according to two metapaths to caputre “publishing habits”:
  - APT: considers ***the topics of their published articles***
  - APV: considers ***the venues in which they select to publish their articles***

- Our used metapath-based similarity measure was the one introduced in [4].



[4] Xiong, Y., Zhu, Y., Yu, P.S.: Top-k similarity join in heterogeneous information networks. IEEE Transactions on Knowledge and Data Engineering 27(6), 1710– 1723 (2015)

A new evaluation framework

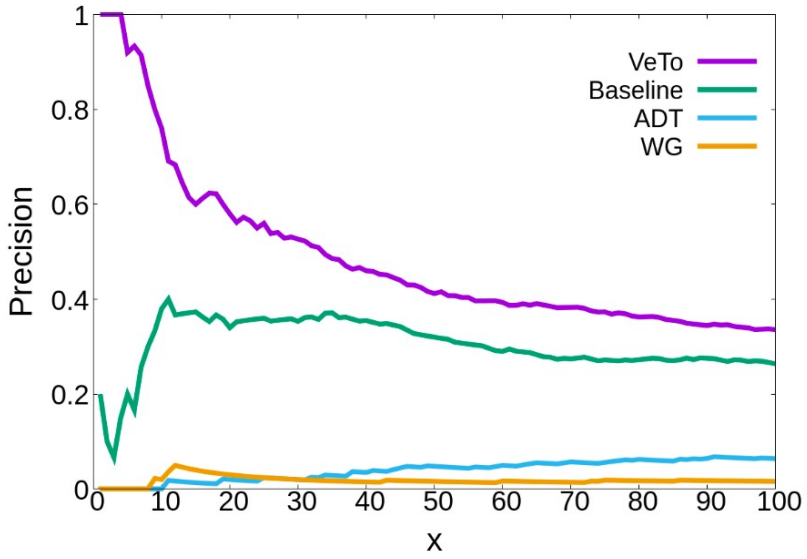
- Building an objective ground truth is almost impossible for many expert finding problems.
- We developed an ***evaluation framework*** to assess the effectiveness of expert set expansion approaches based on a ***fairly objective ground truth***.
- Intuition: gather ***expert lists for real-world applications*** (e.g., PC members, editorial boards) & use them as datasets for ***k-fold cross validation***.
  - Shuffle and split the list E in k disjoint sets E_1, …, E_k
  - Use each E_i as testing set (E^{test}) & the union of the rest sets as training set (E^{train})
    - Training set = set of known experts (E_{kn}) that is given as input to the method
    - Testing set = the ground truth, the "correct" list of new experts (i.e., the expansion set) that we expect in the output
    - Candidates for expansion (set C) are all individuals in the used scholarly knowledge graph.
  - For each E_i examine false & true positives and ***measure precision, recall, F_1-score***. Also ***measure MRR***.

Experimental setup

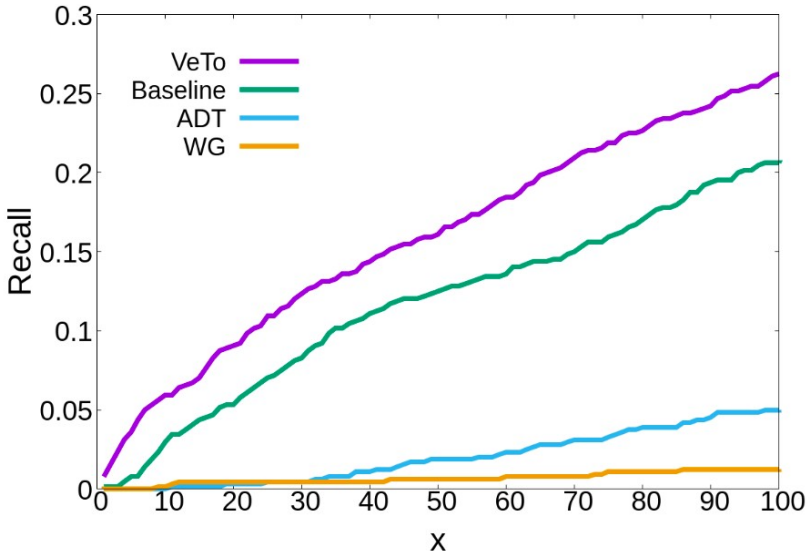
- ***Process:*** based on the proposed framework
- ***Data:***
  - ***Knowledge graph:*** DBLP Scholarly Knowledge Graph (DSKG)
    - Data for ~1.5M academics, their papers between 2000-2017, the corresponding venues and topics.
    - Constructed based on the Aminer's DBLP citation network dataset and topics produced by the CSO Classifier based on the paper abstracts.
  - ***Ground truth:*** PCs of ACM SIGMOD and VLDB conferernces
    - Data gathered by scrapping the official Web pages of the conferences between 2007-2017 & then applying semi-automatic cleaning.
- ***Competitors:***
  - ***Baseline:*** approach that counts the number of papers an academic has published in the corresponding conference, ranks academics based on this number, and provides top ranked academics as the most suitable expansions.
  - ***ADT:*** the best-performing graph-based approach proposed in [5] that attempts to capture the association strength between two academics by considering the paths that relate them to topics based on their papers.
  - ***WG:*** A graph-based approach proposed in [6] which exploits working groups (i.e., co-authors) to capture similarity.

[5] Gollapalli, S.D., Mitra, P., Giles, C.L.: Ranking experts using author-document- topic graphs. In: 13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '13, Indianapolis, IN, USA, July 22 - 26, 2013. pp. 87–96 (2013). https://doi.org/10.1145/2467696.2467707
[6] Balog, K., de Rijke, M.: Finding similar experts. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 821–822. SIGIR '07, Association for Computing Ma- chinery, New York, NY, USA (2007). https://doi.org/10.1145/1277741.1277926, https://doi.org/10.1145/1277741.1277926

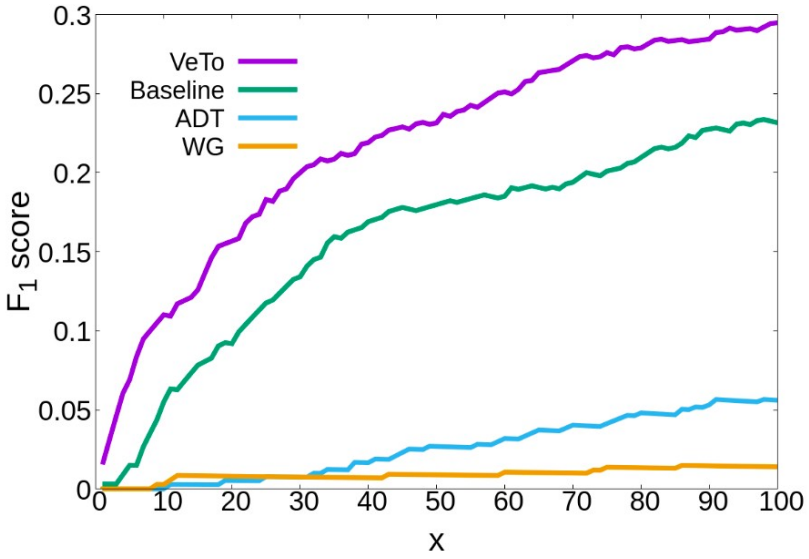
Precision, recall, F1-measure (@x)



(a) Precision



(b) Recall



(c) F_1 score

Fig. 3: Evaluation against competitors for VLDB conference.

MRR

Table 1: MRR based on the folds of each dataset

	Baseline	**ADT**	**WG**	**VeTo**
SIGMOD	0.323	0.043	0.039	0.8
VLDB	0.357	0.046	0.061	1
Total	0.34	0.0445	0.05	0.9

Studying & configuring VeTo

Table 2: MRR of different variants based on the folds of each dataset

	VeTo-APT	**VeTo-APV**	**VeTo**
SIGMOD	0.766	0.766	0.8
VLDB	0.8	0.8	1
Total	0.783	0.783	0.9

Summarizing our contribution

- We introduced VeTo ***a novel approach that effectively deals with the set expansion problem in academia***.
- We proposed ***an evaluation framework that could be used to assess the effectiveness of set expansion approaches*** in a fairly objective way.
- We exploited the developed framework using as expert sets the lists of PCs of known data management conferences to ***evaluate the effectiveness of VeTo against competitors***.
- We provide the expert sets used for our experiments as ***open datasets*** to be used by other researchers.

Thank you!

All flat icons used in the figures have been developed by Freepik or Eucalyp and were downloaded from www