Author Set Identification via Quasi-Clique Discovery

A presentation for DAA 2022/2023

Manuel Gomes

06/01/2023

Table of Contents

Table of Contents

Author Identification

The problem of author identification has been extensively studied, which aims to learn a model to rank potential authors for an anonymous paper based on public information.

Author Identification

The problem of author identification has been extensively studied, which aims to learn a model to rank potential authors for an anonymous paper based on public information.

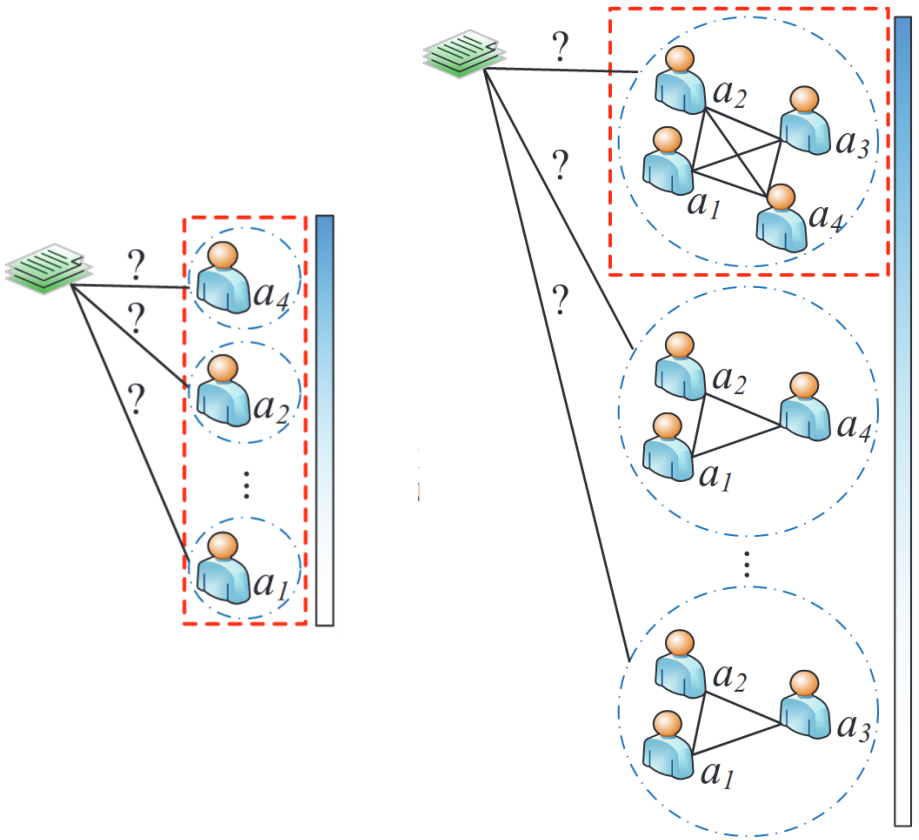
Possible Uses

- Author identification
- Relevance search
- Personalized recommendations
- Reviewer recommendation
- Collaboration discovery

Problem formulation

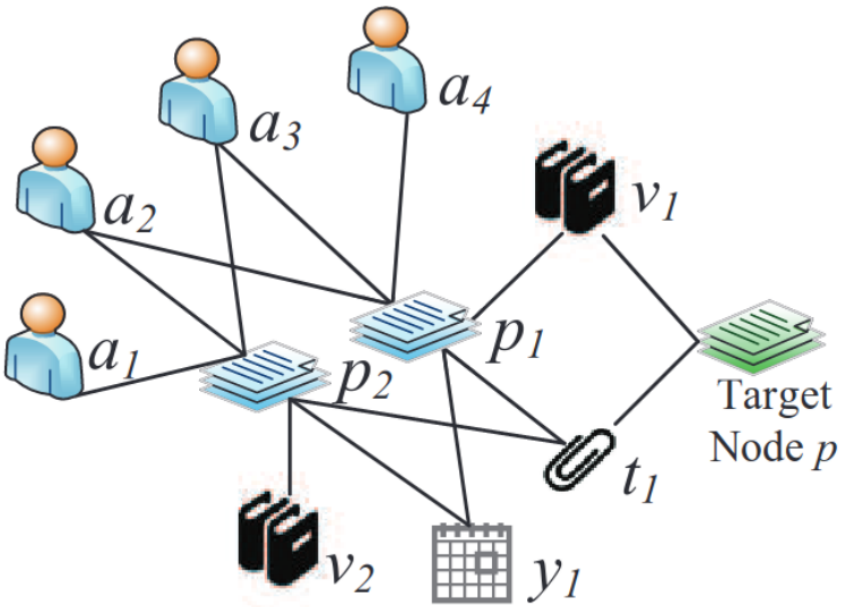
Current approaches:

- Focus on single author identification
- Do not account for the optimal number of writers
- Do not take into account the relation between possible authors



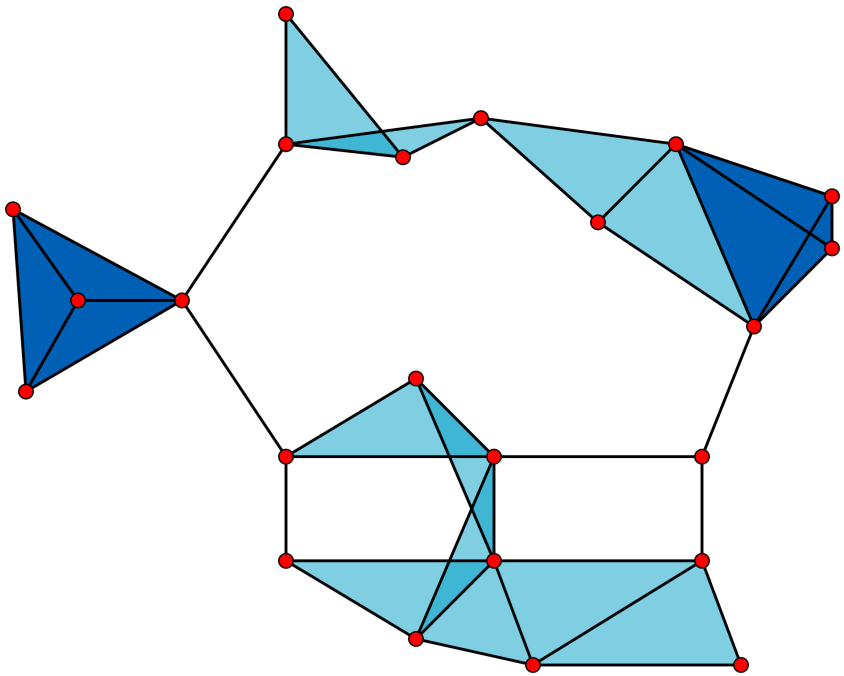
Heterogeneous Information Network (HIN)

- Directed graph
- Nodes and edges have types



Clique (Graph Theory)

- Undirected graph
- Subset of vertices such that every two distinct vertices are adjacent



Quasi-Clique (Graph Theory)

A set of nodes S is an α-quasi-clique if the edge density of the subgraph induced by S exceeds a threshold parameter $\alpha \in (0, 1)$

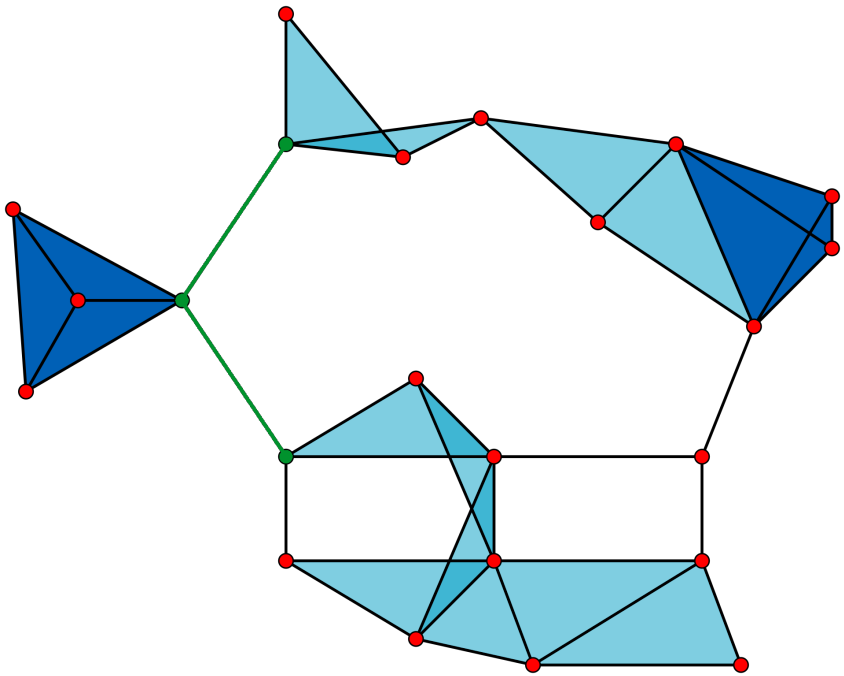
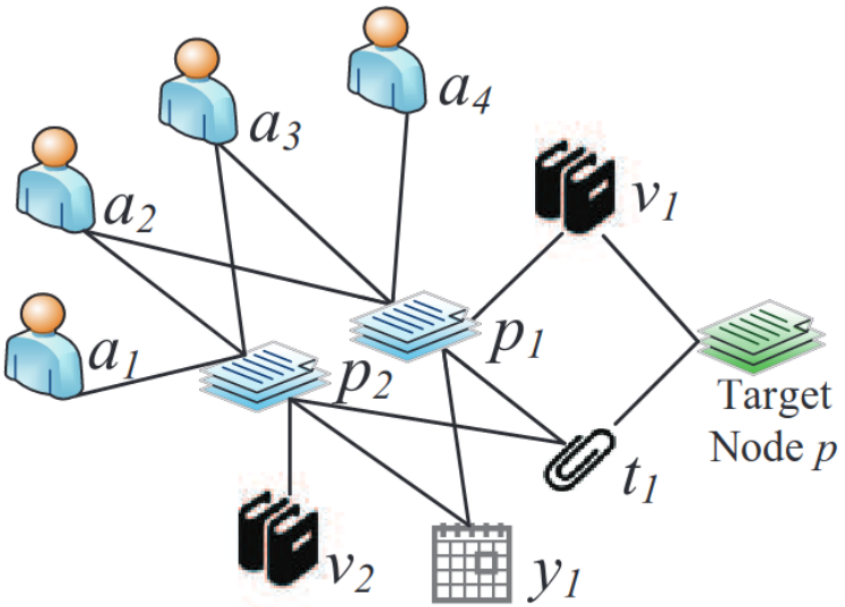


Table of Contents

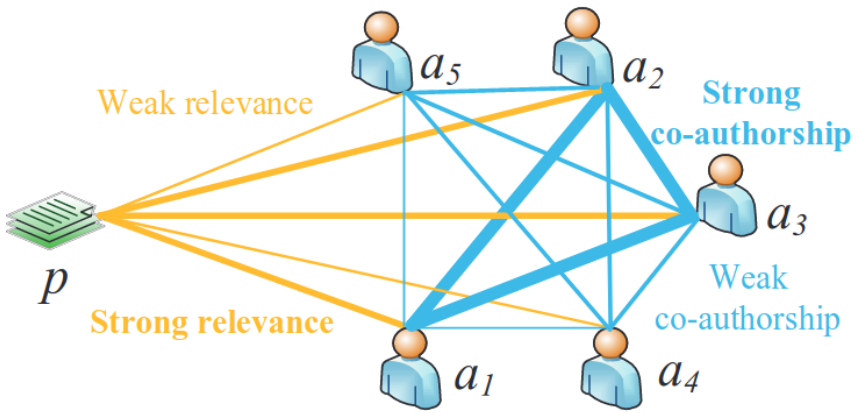
Heterogeneous Information Network (HIN)

- HINs are complex structures
- Complex structures lead to complex algorithms
- How to simply this HIN without removing relevant information?



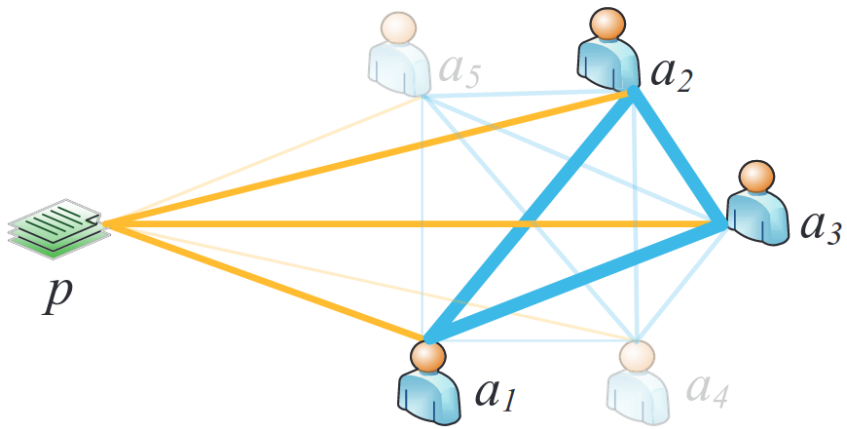
Weighted Paper-Ego-Network

- Extraction of a task-guided embedding to learn the low-dimensional representation of a network
- Creation of a simpler weighted graph



Optimal Author Subset

- To find an optimal author subset that is related to the anonymous paper is an NP-hard problem
- Discovering the largest clique is inapproachable and the clique concept is in practice too strict to miss a single edge in an otherwise dense subgraph
- Quasi-clique has been significantly used to discover dense subgraphs
- Maximum quasi-clique discovery is also an NP-hard problem



Algorithm description

- Optimization problem
- Local search algorithm
- Density calculation

Algorithm 1: OQCCE

Input : Weighted paper ego network $G_p = (V, E, W)$; maximum number of iterations I_{max}; the constrained node p

Output: A subset of nodes $S \subseteq V$ and $p \in S$

$S \leftarrow p$, $b_1 \leftarrow$TRUE, $i \leftarrow 1$;

while b_1 *and* $i \leq I_{max}$ **do**

  $b_2 \leftarrow$TRUE;

  while b_2 **do**

    if *there exists* $u \in V \setminus S$ *and* $g_{\alpha,\beta}(S \cup \{u\}) \geq g_{\alpha,\beta}(S)$ **then**

      $S \leftarrow S \cup \{u\}$;

    else

      $b_2 \leftarrow$FALSE;

  if *there exists* $u \in S$ *and* $u \neq p$ *and* $g_{\alpha,\beta}(S \setminus \{u\}) \geq g_{\alpha,\beta}(S)$ **then**

    $S \leftarrow S \setminus \{u\}$;

  else

    $b_1 \leftarrow$FALSE;

  $i \leftarrow i + 1$;

Table of Contents

Datasets Used

- Large scale datasets

Dataset	# papers	# authors	# terms	# venues
AMiner-I	8821	12660	12467	5
AMiner-II	35349	36247	31446	14

Evaluation Metrics

- $P = \frac{|S'_a \cap S_a|}{|S'_a|}$
  - Not sensitive to false negatives
- $R = \frac{|S'_a \cap S_a|}{|S_a|}$
  - Not sensitive to false positives
- $F1 = \frac{2*P*R}{P+R}$

- S_a - real author set
- S'_a - returned author set

Dataset	# papers	# authors	# terms	# venues
AMiner-I	8821	12660	12467	5
AMiner-II	35349	36247	31446	14

Results

Methods			Evaluation						Avg.
			P (↑)	R (↑)	J (↑)	F1 (↑)	MAP (↑)	RMSE (↓)	
Top-5	Similarity measure	PTPA	0.2716 (2)	0.5007 (7)	0.2310 (2)	0.3356 (2)	**0.6109** (1)	0.1714 (2)	2.67
		PCPA	0.2098 (7)	0.3937 (11)	0.1680 (7)	0.2614 (7)	0.4718 (9)	0.1714 (2)	7.16
	Feature method	LR	0.2160 (5)	0.3915 (12)	0.1827 (6)	0.2657 (4)	0.4834 (7)	0.1714 (2)	6.00
		SVM	0.2493 (3)	0.4562 (9)	0.2154 (4)	0.3081 (3)	0.5451 (3)	0.1714 (2)	4.00
		Bayesian	0.2209 (4)	0.4075 (10)	0.1888 (5)	0.2733 (5)	0.4951 (6)	0.1714 (2)	5.33
	HetNetE		0.2123 (6)	0.3870 (13)	0.1669 (8)	0.2616 (6)	0.4571 (11)	0.1714 (2)	7.66
Top-10	Similarity measure	PTPA	0.1555 (9)	0.5779 (2)	0.1454 (10)	0.2365 (9)	0.5897 (2)	0.5023 (3)	5.83
		PCPA	0.1388 (11)	0.5066 (5)	0.1257 (13)	0.2110 (11)	0.4517 (12)	0.5023 (3)	9.10
	Feature method	LR	0.1358 (13)	0.5005 (8)	0.1270 (12)	0.2059 (13)	0.4664 (10)	0.5023 (3)	9.83
		SVM	0.1629 (8)	**0.5988** (1)	0.1538 (9)	0.2477 (8)	0.5296 (4)	0.5023 (3)	5.50
		Bayesian	0.1364 (12)	0.5010 (6)	0.1277 (11)	0.2069 (12)	0.4767 (8)	0.5023 (3)	8.67
	HetNetE		0.1506 (10)	0.5347 (3)	0.2269 (3)	0.2275 (10)	0.4435 (13)	0.5023 (3)	7.00
ASI			**0.4589** (1)	0.5284 (4)	**0.4009** (1)	**0.4712** (1)	0.5295 (5)	**0.1123** (1)	2.00

Methods			Evaluation						Avg.
			P (↑)	R (↑)	J (↑)	F1 (↑)	MAP (↑)	RMSE (↓)	
Top-5	Similarity measure	PTPA	0.3391 (2)	0.5899 (6)	0.2886 (2)	0.4108 (2)	0.7165 (3)	0.2880 (2)	2.83
		PCPA	0.3287 (3)	0.5743 (8)	0.2776 (4)	0.3986 (3)	0.6595 (6)	0.2880 (2)	4.33
	Feature method	LR	0.3113 (4)	0.5400 (9)	0.2645 (5)	0.3769 (4)	0.6605 (5)	0.2880 (2)	4.83
		SVM	0.2202 (7)	0.4553 (12)	0.1674 (11)	0.2803 (9)	**0.9948** (1)	0.2880 (2)	7.00
		Bayesian	0.2964 (5)	0.5144 (10)	0.2491 (6)	0.3587 (5)	0.6458 (8)	0.2880 (2)	6.00
	HetNetE		0.2645 (6)	0.4561 (11)	0.2078 (7)	0.3191 (6)	0.6021 (12)	0.2880 (2)	7.33
Top-10	Similarity measure	PTPA	0.1927 (8)	**0.6624** (1)	0.1795 (8)	0.2884 (7)	0.6913 (4)	0.8536 (3)	5.16
		PCPA	0.1913 (9)	0.6531 (2)	0.1778 (9)	0.2860 (8)	0.6363 (10)	0.8536 (3)	6.83
	Feature method	LR	0.1857 (10)	0.5779 (7)	0.1729 (10)	0.2775 (10)	0.6382 (9)	0.8536 (3)	8.16
		SVM	0.1101 (13)	0.4553 (12)	0.0943 (13)	0.1702 (13)	0.9948 (1)	0.8536 (3)	5.00
		Bayesian	0.1786 (11)	0.6157 (4)	0.1661 (12)	0.2673 (11)	0.6227 (11)	0.8536 (3)	8.66
	HetNetE		0.1720 (12)	0.6350 (3)	0.2858 (3)	0.2564 (12)	0.5602 (13)	0.8536 (3)	7.66
ASI			**0.5981** (1)	0.6019 (5)	**0.4943** (1)	**0.5720** (1)	0.6566 (7)	**0.2058** (1)	2.66

Table of Contents

Conclusion

- Author identification is a very useful (and interesting) subject matter
- Simplification of HIN to solve a maximum quasi-clique problem is very ingenious
- Author set identification is a definite improvement over state of the art