Author Name Disambiguation using Markov Chain Monte Carlo (MCMC)

Master's Thesis

Nagaraj Bahubali Asundi

First supervisor: Dr. Zeyd Boukhers
Second supervisor: Dr. Claudia Schon

Institute for Web Science and Technologies (WeST)
University of Koblenz-Landau

January 2, 2023

Outline

Table of Contents

Introduction

Author Name Disambiguation (AND)

- **Citation:** a reference string that contains bibliographic information about a scientific paper
- **Aim:** map author name in citations to a real-world person

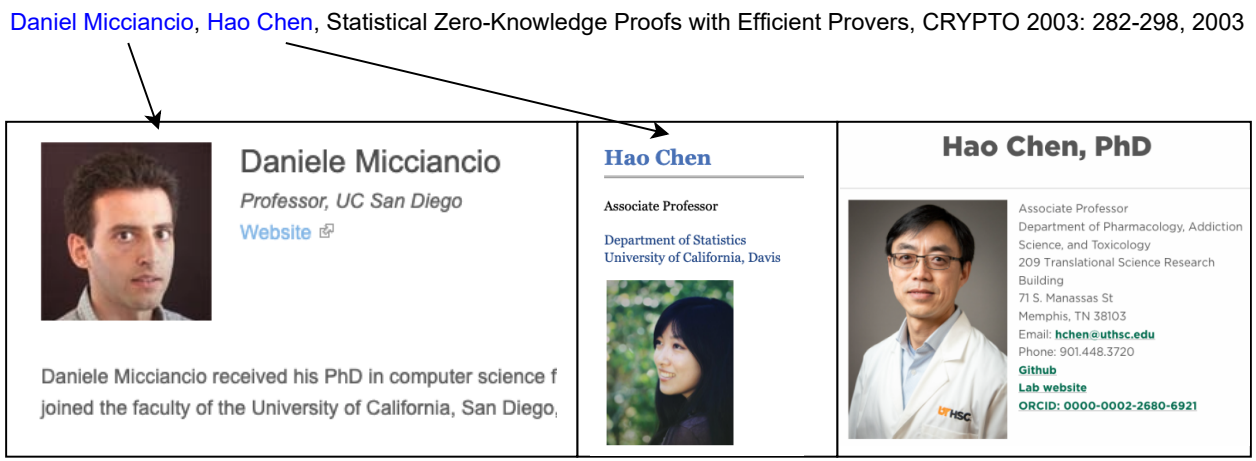


Table of Contents

Motivation

Challenges

- **Homonymy:** authors sharing the same names
  - Hao Chen, Associate Professor from California
  - Hao Chen, Associate Professor from Memphis
- **Synonymy:** author sharing different name variants
  - JangMyung Lee, Pusan National University, South Korea
  - J. Lee, Pusan National University, South Korea

Motivation

- Quality of scientific data gathering
- Incorrect identification and credit attribution to authors

Table of Contents

Problem Statement and Contributions 1/2

Problem Statement: Process a set of citations containing authors and their associated papers into clusters, each containing a set of papers pertaining to a real-world author.

- Let $P = \{p_1, p_2, ..., p_n\}$ be a set of n papers
- Let $A = \{a_1, a_2, ..., a_m\}$ be a set of m real-world authors
- A paper p_i belonging to real-world author a_j is denoted by (p_i, a_j)
- **Goal:** partition P into m disjoint clusters $C_1, C_2, ..., C_m$ such that $C_1 = \{p_i | (p_i, a_1)\}$, ..., $C_m = \{p_i | (p_i, a_m)\}$

Problem Statement and Contributions 2/2

Contributions

- Introduces *AND-MCGC* - Author Name Disambiguation using Markov Chain-based Graph Clustering
- Employs multiple factors to achieve AND: research area, publication pattern over the years, patterns of co-authorship, and overlap of affiliations
- Conducts extensive experiments to analyze the contribution of each factor
- Provides an open-source implementation of the proposed approach[1]

[1] https://github.com/nagaraj-bahubali/author-name-disambiguation-using-mcmc

Table of Contents

Overview

1. Parse raw data and form citations

   author_id | paper_id | author_name <> co_author_1@affiliation_1; co_author_2@affiliation_2;...;co_author_n@affiliation_n <> title <> venue <> year

2. Organize citations according to namespaces - atomic names
   - Chen Li and Cheng Li $\rightarrow$ C Li
   - Steven Smith and Steffen Smith $\rightarrow$ S Smith

3. Construct graph

4. Perform graph actions

Parsed Citations

Bing Li

P_1 : Bing Li@Uni-Trier;Fangshi Wang@Uni Trier<>Anisotropic Convolution for Image Classification<>CBMI<>2017

P_2 : Bing Li@uni-trier;Zhipeng Zhang@Uni Uulm<>Object Tracking via Attention Retrieval<>AIPR<>2004

Da Xing

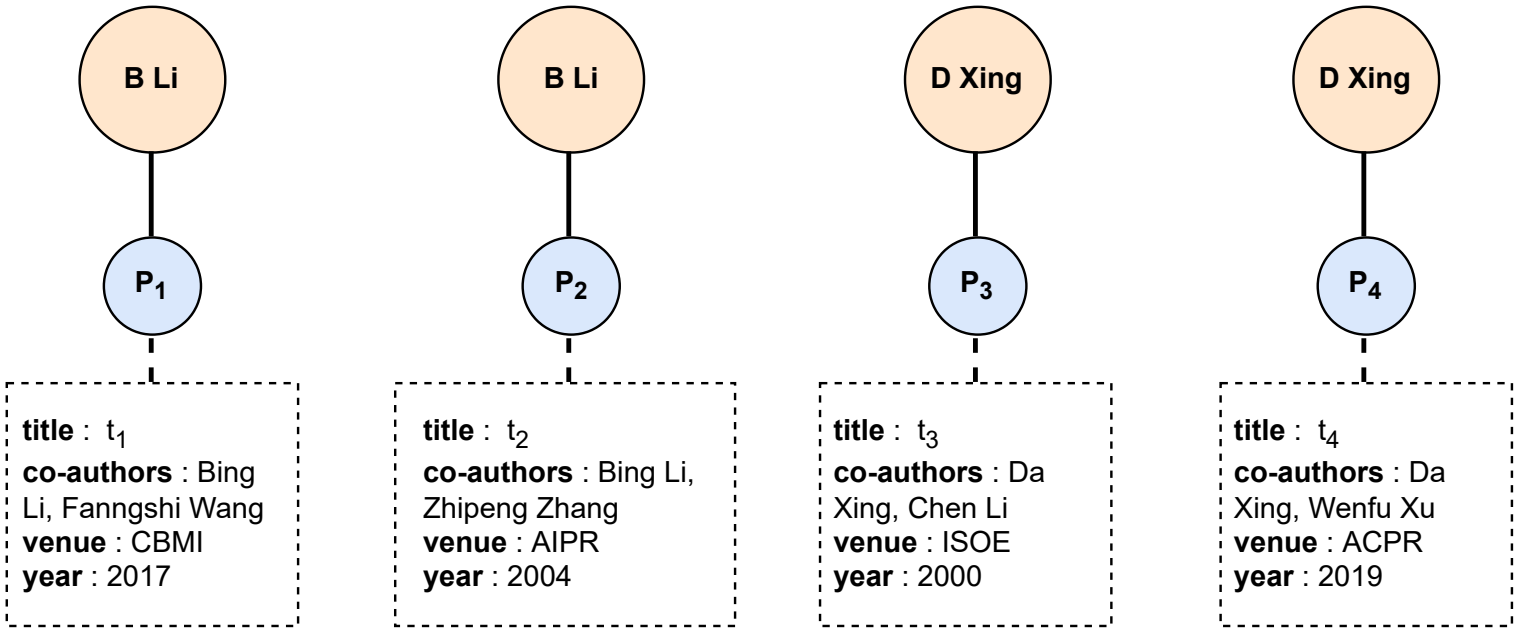
P_3 : Da Xing@South China Uni;Chen Li@South China Uni<>Monte Carlo Study of Light<>ISOE<>2000

P_4 : Da Xing@Fudan Uni;Wenfu Xu@Lanzhou Uni<>Algorithm for robot path planning<>ACPR<>2019

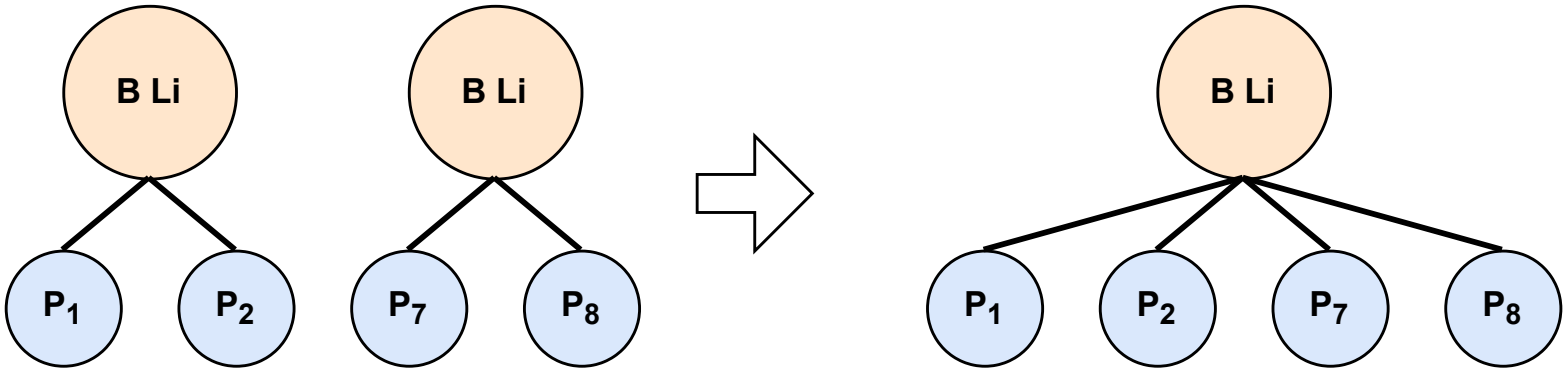
Name Spaces

B Li, D Xing

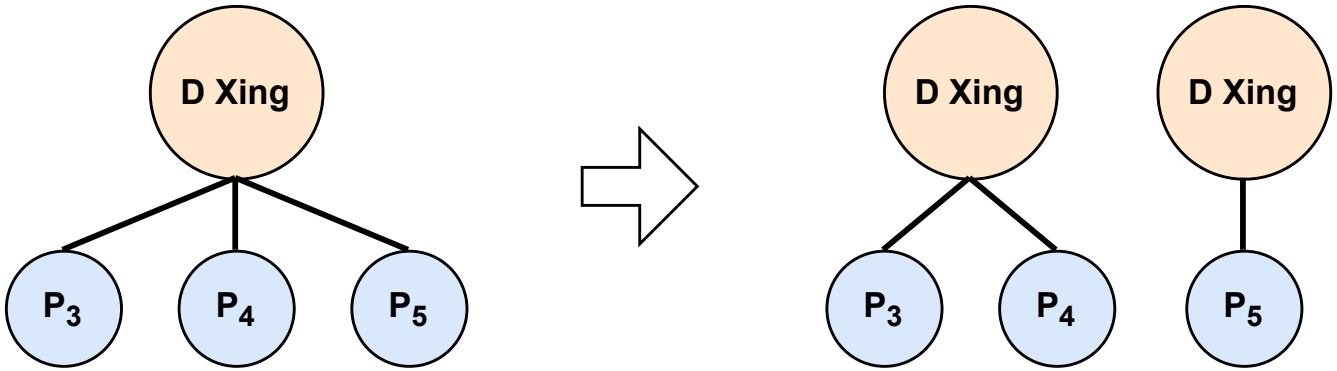
Graph Construction



Graph Actions

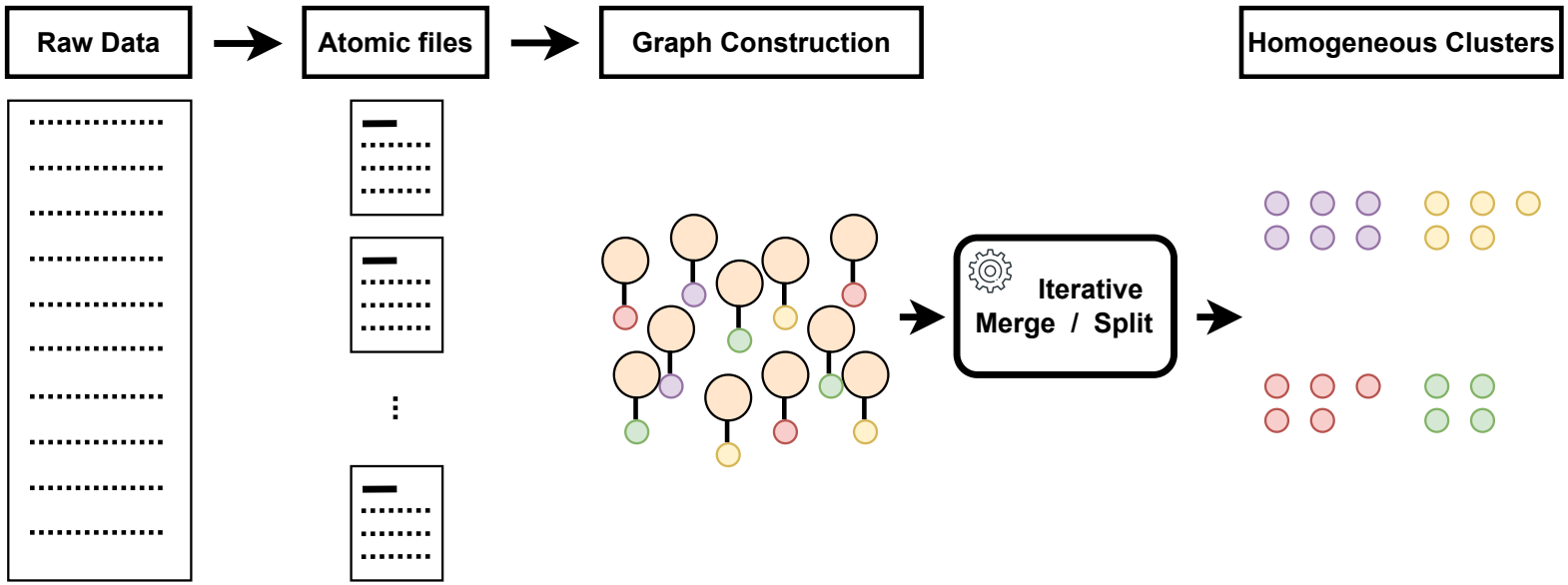


(a) Merging the "Bing Li" papers.



(b) Splitting the "Da Xing" papers.

AND-MCGC - Architecture



MCMC in the context of AND

Metropolis-Hastings Algorithm

- Used for estimating target distribution
- Start with an initial random variable and propose subsequent variables in the form of a Markov chain
- Random variables are accepted/ rejected using acceptance criteria
- Propose enough times so that sampling distribution reflects target distribution

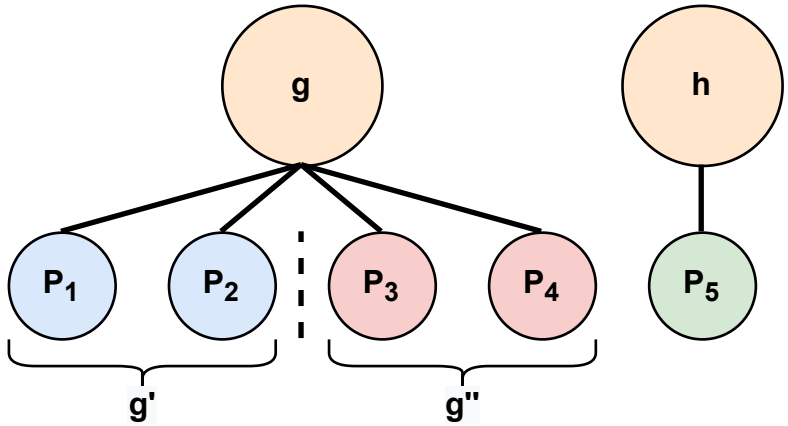
AND-MCGC

- target distribution $\rightarrow$ *ideal state* of the graph
- Start with an initial state $\rightarrow$ no. of graphlets $=$ no. of citations
- Propose certain changes in the topology of the graph (merge/ split)
- accept/ reject the proposals using acceptance criteria
- Propose enough times to reach *ideal state*

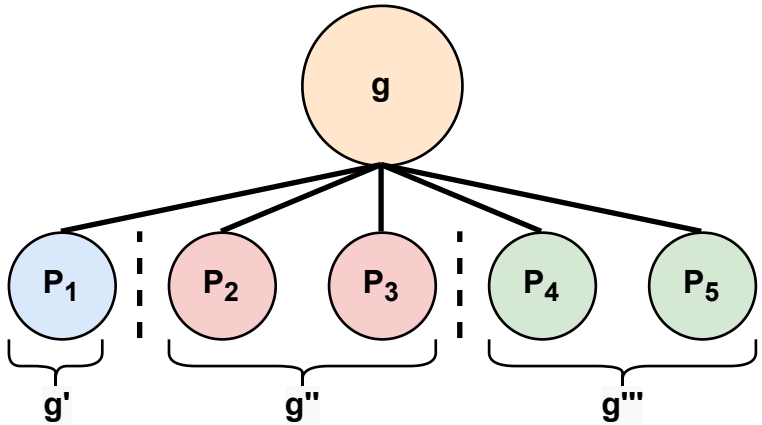
Acceptance terms 1/2

1. Domain Similarity - α

- $\alpha_{merge} = \frac{\text{Sim b/w papers in } g \text{ \& } h}{\text{Sim b/w papers in } g' \text{ \& } g''}$

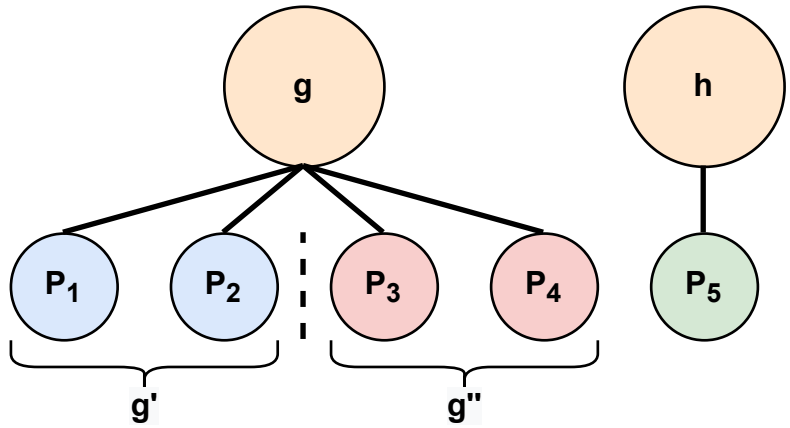


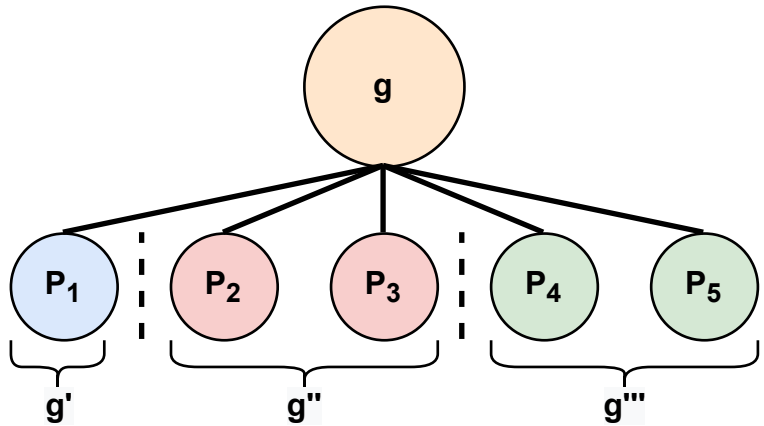
- $\alpha_{split} = \frac{\text{Sim b/w papers in } g' \text{ \& } g''}{\text{Sim b/w papers in } g' \cup g'' \text{ \& } g'''}$



Acceptance terms 2/2

2. Co-authorship Overlap - β
   - Jaccard similarity of the co-authors
3. Publication Pattern - γ
   - Distribution of topics over time
   - likelihood b/w graphlets
4. Affiliation Overlap - κ
   - Jaccard similarity of the co-authors' affiliations





acceptance ratio $= \alpha * \beta * \gamma * \kappa$

Table of Contents

Experimental setup

Dataset

- Aminer-534K[2]
- 35,107 citations
- 6,399 unique real-world authors

Latent Topic Inference

- Gaussian LDA by Das et al. [1]
- number of topics K = 15
- number of iterations I = 20

[2]https://static.aminer.cn/misc/na-data-kdd18.zip

Table of Contents

Evaluation

- pairwise measures by Levin et al. [2]

Baseline Comparison

Model	Precision	Recall	F1
GHOST [3]	81.62	40.43	50.23
Louppe et al. [4]	57.09	77.22	63.10
Zhang et al. [5]	70.63	59.53	62.81
Chen et al. [6]	65.59	69.96	65.71
Pooja et al. [7]	62.60	76.10	66.90
Aminer-18 [8]	77.96	63.03	67.79
Wang et al. [9]	82.23	67.23	72.92
LAND [10]	77.24	61.21	64.18
AND-MCGC$_{\alpha}$	47.07	54.44	43.71
AND-MCGC$_{\gamma}$	53.16	56.71	46.76
AND-MCGC$_{all}$	51.73	67.67	50.79

Table of Contents

Future Enhancements

- Facilitate the integration process with global topic modeling
- Use abstracts/whole text of papers to extract latent topics
- Take advantage of the venues by profiling them according to the domain of the papers

Thank You

References I

[1] R. Das, M. Zaheer, and C. Dyer, "Gaussian lda for topic models with word embeddings," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 795–804.

[2] M. Levin, S. Krawczyk, S. Bethard, and D. Jurafsky, "Citation-based bootstrapping for large-scale author disambiguation," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 5, pp. 1030–1047, 2012.

[3] X. Fan, J. Wang, X. Pu, L. Zhou, and B. Lv, "On graph-based name disambiguation," *Journal of Data and Information Quality (JDIQ)*, vol. 2, no. 2, pp. 1–23, 2011.

References II

[4] G. Louppe, H. T. Al-Natsheh, M. Susik, and E. J. Maguire, "Ethnicity sensitive author disambiguation using semi-supervised learning," in *international conference on knowledge engineering and the semantic web*, Springer, 2016, pp. 272–287.

[5] B. Zhang and M. Al Hasan, "Name disambiguation in anonymized graphs using network embedding," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1239–1248.

[6] Y. Chen, H. Yuan, T. Liu, and N. Ding, "Name disambiguation based on graph convolutional network," *Scientific Programming*, vol. 2021, 2021.

[7] K. Pooja, S. Mondal, and J. Chandra, "Exploiting similarities across multiple dimensions for author name disambiguation," *Scientometrics*, vol. 126, no. 9, pp. 7525–7560, 2021.

References III

[8] Y. Zhang, F. Zhang, P. Yao, and J. Tang, "Name disambiguation in aminer: Clustering, maintenance, and human in the loop.," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1002–1011.

[9] H. Wang, R. Wan, C. Wen, *et al.*, "Author name disambiguation on heterogeneous information network with adversarial representation learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 238–245.

[10] C. Santini, G. A. Gesese, S. Peroni, A. Gangemi, H. Sack, and M. Alam, "A knowledge graph embeddings based approach for author name disambiguation using literals," *arXiv preprint arXiv:2201.09555*, 2022.