A Comparative Data/Hypernetwork Analysis of Publication and Citation Profiles from Dimensions Datasets

Moses Boudourides

Master's in Data Science Online Program
School of Professional Studies
Northwestern University

Moses.Boudourides@northwestern.com

Seminar at the Lehrstuhl für Human-Centered Computing
School of Social Sciences and Technology
Technische Universität München
Tuesday, October 29, 2024

- The analyzed dataset of publications was compiled from the Dimensions.ai database through two search stages:

  (i) retrieving all publications authored by each scholar,
  (ii) retrieving all publications citing those in step (i), and
  (iii) retrieving publications with grant funding from those identified in step (i).

- Publication A is said to "cite" publication B if B appears in the reference list of A. In this case, A is referred to as a "citation" of B.

- The following 16 fields (attributes) of publications and grants were retrived: `id`, `authors`, `title`, `date`, `doi`, `type`, `reference_ids`, `category_for`, `concepts`, `concepts_scores`, `times_cited`, `altmetric`, `field_citation_ratio`, `relative_citation_ratio`, `grants` and `funding _usd`. Most of these fields are self-explanatory; however, they will be discussed as they are encountered to clarify their significance and what they reveal about the dataset. Documentation for all the Dimensions.ai fields can be found in `https://docs.dimensions.ai/dsl/2.0.0/datasource-publications.html`.

Comparing Six Authors of Publications on Data Science

- **Marinka Zitnik**, Assistant Professor, Department of Biomedical Informatics, Harvard Medical School.
- **Utku Pamuksuz**, Associate Clinical Professor, Data Science Institute, University of Chicago.
- **Kyriaki Kalimeri**, Researcher, ISI Foundation, Turin, Italy, and Senior Research Consultant at UNICEF.
- **Jason Rute**, Postdoctoral Researcher, MIT-IBM Watson AI Lab.
- **Jesse Lecy**, Associate Professor, Center on Technology, Data and Society, School of Public Affairs, Arizona State University.
- **Carlos Fernandez-Granda**, Associate Professor of Mathematics and Data Science, Courant Institute of Mathematical Science and Center for Data Science, New York University.

After removing duplicate publications, the unique count of each publication can be established by enumerating the distinct `ids` associated with that publication. It is important to note that if a publication was retrieved under two or more different `types`—for example, as both an `article` (or `proceeding` or `chapter`) and a `preprint`—the `preprint` version was excluded. This ensures that all publications in our dataset are uniquely characterized by their `ids`.

Scholars	Publications	Citations	Citations (D)	Funded Publications
Zitnik	166	166	6837	70
Pamuksuz	6	62	64	1
Kalimeri	78	78	816	6
Rute	21	261	275	6
Lecy	44	1326	1483	3
Fernandez-Granda	88	2073	2585	52

Yearly Densities of Publications

If one has a time series of counts of certain entities $\mathbf{x} = (x(t_1), x(t_2), \ldots, x(t_m))$, where the timestamps $t_1, t_2, \ldots, t_m$ are sorted in increasing order, the *time density* of $\mathbf{x}$ is typically defined as the average value of $\mathbf{x}$ over the number of dates, given by

$$\frac{1}{m} \sum_{i=1}^{m} x(t_i).$$

Since each $x(t_i)$ represents a count, it is a nonnegative integer. To exclude trivial contributions from zero counts, in other words, by limiting the data on *active time*, the *active density* of the $x(t_i)$ can be defined as

$$\frac{1}{p} \sum_{k=1}^{p} x(t_{n_k}),$$

where $t_{n_1}, t_{n_2}, \ldots, t_{n_p}$ is a subsequence of $t_1, t_2, \ldots, t_m$ such that all counts in this subsequence are positive ($x(t_{n_k}) > 0$ for all $k = 1, 2, \ldots, p$).

Scholars	Publications	No. of Actives Years	Active Years	Yearly Publication Density
Zitnik	166	13	2012-2024	12.769231
Pamuksuz	6	5	2012, 2016, 2019, 2021, 2024	0.461538
Kalimeri	78	14	2010-2014, 2016-2024	5.200000
Rute	21	11	2011-2012, 2014-2021, 2024	1.500000
Lecy	44	15	2010-2024	2.933333
Fernandez-Granda	88	12	2012-2013, 2015-2024	6.769231

Scholars	Citations	No. of Actives Years	Actives Years	Yearly Citation Density
Zitnik	166	13	2012-2024	12.769231
Pamuksuz	62	8	2012, 2016, 2019-2024	4.769231
Kalimeri	78	14	2010-2014, 2016-2024	5.200000
Rute	261	14	2011-2024	18.642857
Lecy	1326	15	2011-2025	88.400000
Fernandez-Granda	2073	14	2012-2025	148.071429

Scholars	Funded Publications	Actives Years	Actives Years	Yearly Funded Publications Density
Zitnik	70	12	2013-2024	5.833333
Pamuksuz	1	1	2012	1.000000
Kalimeri	6	4	2016, 2018, 2020, 2022	0.857143
Rute	6	5	2012, 2014, 2016-2017, 2019	0.750000
Lecy	3	3	2013, 2017, 2021	0.333333
Fernandez-Granda	52	9	2012, 2017-2024	4.000000

Statistics of Numerical Fields (Attributes of Publications

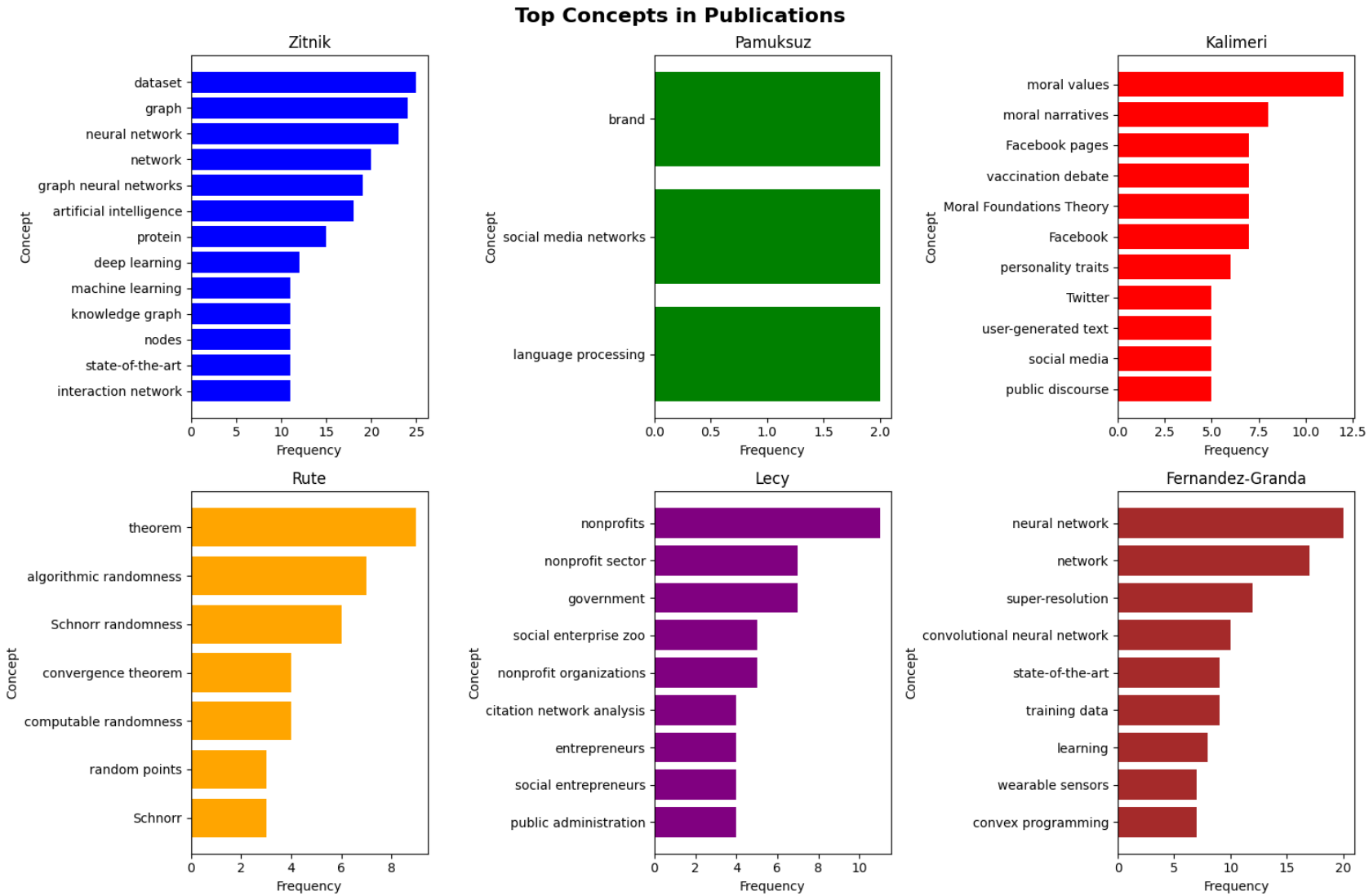
Scholars	Funded Publications	total_funding_usd mean	total_funding_usd std	total_funding_usd N
Zitnik	70	25152600.100000	48567316.6	70
Pamuksuz	1	759000.000000	nan	1
Kalimeri	6	3463000.200000	1960596.1	6
Rute	6	2810872.200000	4398842.8	6
Lecy	3	48191535.700000	71989270.8	3
Fernandez-Granda	52	71000865.000000	291004069.9	52

Scholars	Publications	altmetric	field_citation_ratio	relative_citation_ratio
Zitnik	166	40.0 (86.6), N=123	12.5 (29.9), N=93	3.3 (5.0), N=42
Pamuksuz	6	1.8 (1.5), N=4	5.0 (4.8), N=5	N/A
Kalimeri	78	17.0 (39.5), N=39	4.7 (8.6), N=51	0.6 (0.3), N=5
Rute	21	42.5 (101.8), N=10	12.4 (43.4), N=17	N/A
Lecy	44	30.1 (85.0), N=22	12.9 (20.0), N=38	1.4 (1.4), N=6
Fernandez-Granda	88	29.5 (109.9), N=53	13.0 (49.6), N=63	2.3 (2.2), N=9

Scholars	Citations	altmetric	field_citation_ratio	relative_citation_ratio
Zitnik	0	nan (nan), N=0	nan (nan), N=0	nan (nan), N=0
Pamuksuz	62	3.7 (3.5), N=22	7.1 (11.7), N=27	1.3 (1.3), N=2
Kalimeri	0	nan (nan), N=0	nan (nan), N=0	nan (nan), N=0
Rute	261	21.2 (98.9), N=121	8.2 (15.5), N=147	1.8 (1.7), N=16
Lecy	1326	15.0 (43.4), N=732	10.1 (39.0), N=905	5.1 (17.3), N=132
Fernandez-Granda	2073	17.4 (82.7), N=972	9.8 (49.0), N=946	1.8 (1.9), N=111

Scholars	Funded Publications	altmetric	field_citation_ratio	relative_citation_ratio
Zitnik	70	57.7 (112.7), N=63	20.9 (38.4), N=49	3.7 (5.3), N=37
Pamuksuz	1	nan (nan), N=0	0.5 (nan), N=1	N/A
Kalimeri	6	15.8 (26.3), N=4	6.6 (5.7), N=5	0.9 (nan), N=1
Rute	6	82.2 (162.5), N=4	39.9 (78.6), N=5	N/A
Lecy	3	43.3 (69.0), N=3	4.8 (0.9), N=2	2.5 (2.2), N=2
Fernandez-Granda	52	48.9 (144.1), N=30	6.8 (12.2), N=40	2.3 (2.2), N=9

Top Concepts



Top Research Areas

Citation Graphs

- A "citation graph" (or "citation network") is a directed graph where vertices represent publications and edges denote citation relationships between them. Specifically, if publication v cites publication u, an edge is drawn from vertex u to vertex v in the citation graph.
- More precisely, the above defines a "publications citation graph" (i.e., a citation network among publications). Alternatively, one could define an "authors' citation graph," wherevertices represent authors and the edges represent citation relationships between them.
- In particular, a "local citation graph" (or "local citation network") of publications/authors is defined by one or more focal publications/authors, their citations and all existing citations among the publications/authors that reference them.
- A local citation network, as defined here, constitutes an *ego–centered network*, where the focal publication(s)/author(s) serve as the *ego*(s) and their citations represent the *alters*. It is important to note that since citation graphs are directed acyclic graphs (DAGs), all links connected to the focal publication(s)/author(s) (whereby a citation constitutes an outgoing link) and among their citations are inherently non-reciprocal.

Three measures of citation graphs

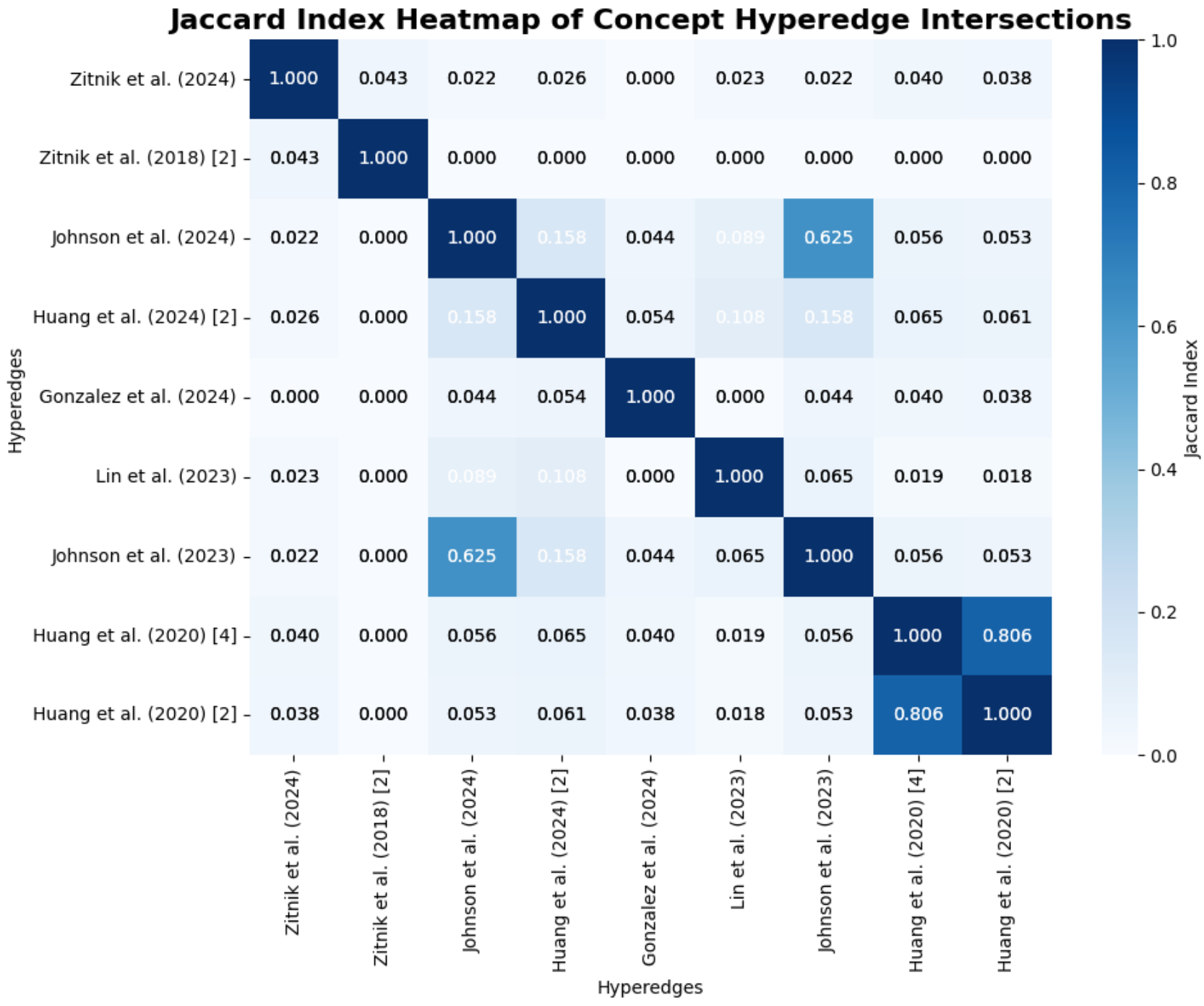
- "Self–citations" refer to instances where a publication cites a previously published work that shares at least one co-author with the citing publication.
- The "clustering coefficient" is a measure used in graph theory and network science to quantify the degree to which vertices in a graph tend to cluster together. For a given vertex, it assesses how many of its neighbors are also neighbors with each other, forming triangles.
- Ronald Burt's concept of "node constraint" quantifies how limited a node's structural autonomy is within its local network. Specifically, it measures the extent to which a node's connections are concentrated among a few closely-knit neighbors, capturing the idea of redundancy in the pattern of network connections.

Scholars	Self–Citations	Average Clustering Coefficient	Average Node Constraint
Zitnik	233	0.070271	NaN
Pamuksuz	1	0.009359	0.055586
Kalimeri	107	0.208957	NaN
Rute	3	0.018827	0.370378
Lecy	18	0.018724	0.176018
Fernandez-Granda	37	0.110834	0.232471

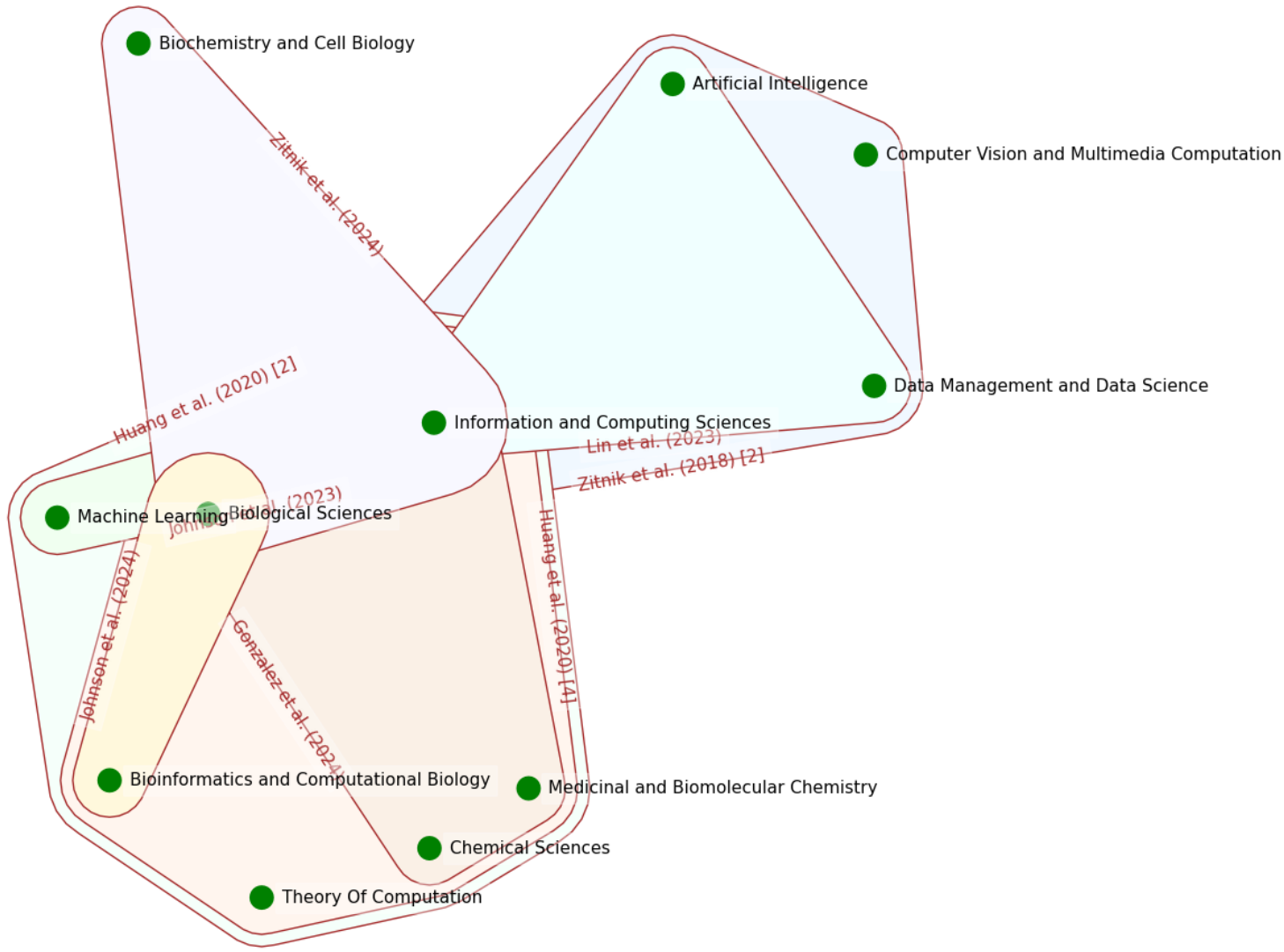
Semantic/Structural Alignment between Publications and Citations

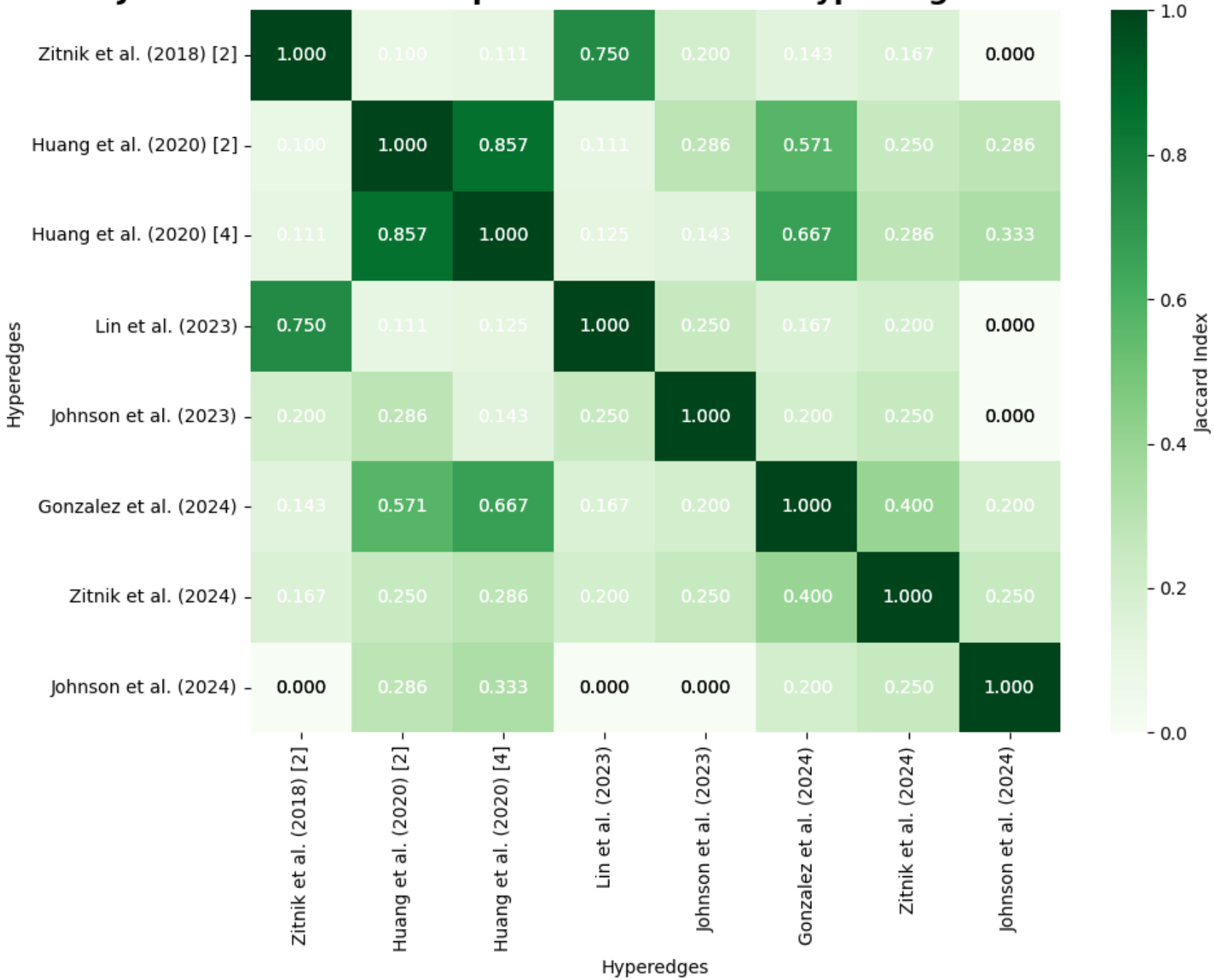
- Setting a threshold value of 0.5, the concepts/research areas from publications that exceeded this threshold were selected for analysis.
- For the citations, those with an in-degree (or out-degree) greater than 5 were selected from the local citation network. Additionally, a threshold value of 0.5 was applied to the concepts/research areas of these citations to isolate their most relevant concepts.
- To compare the collected sets of concepts/research areas, a hypergraph was constructed in which the hyperedges represent individual publications (from both a publication and its citations), while the nodes correspond to the concepts/research areas present in each publication.
- Two visualization techniques were applied: (i) a plot of the hypergraph displaying publications as hyperedges and concepts/research areas as nodes, and (ii) the computation of Jaccard indices to quantify the similarity between pairs of hyperedges/publications.

The Hypergraph of Concepts of Marinka Zitnik in 9 Top Citations

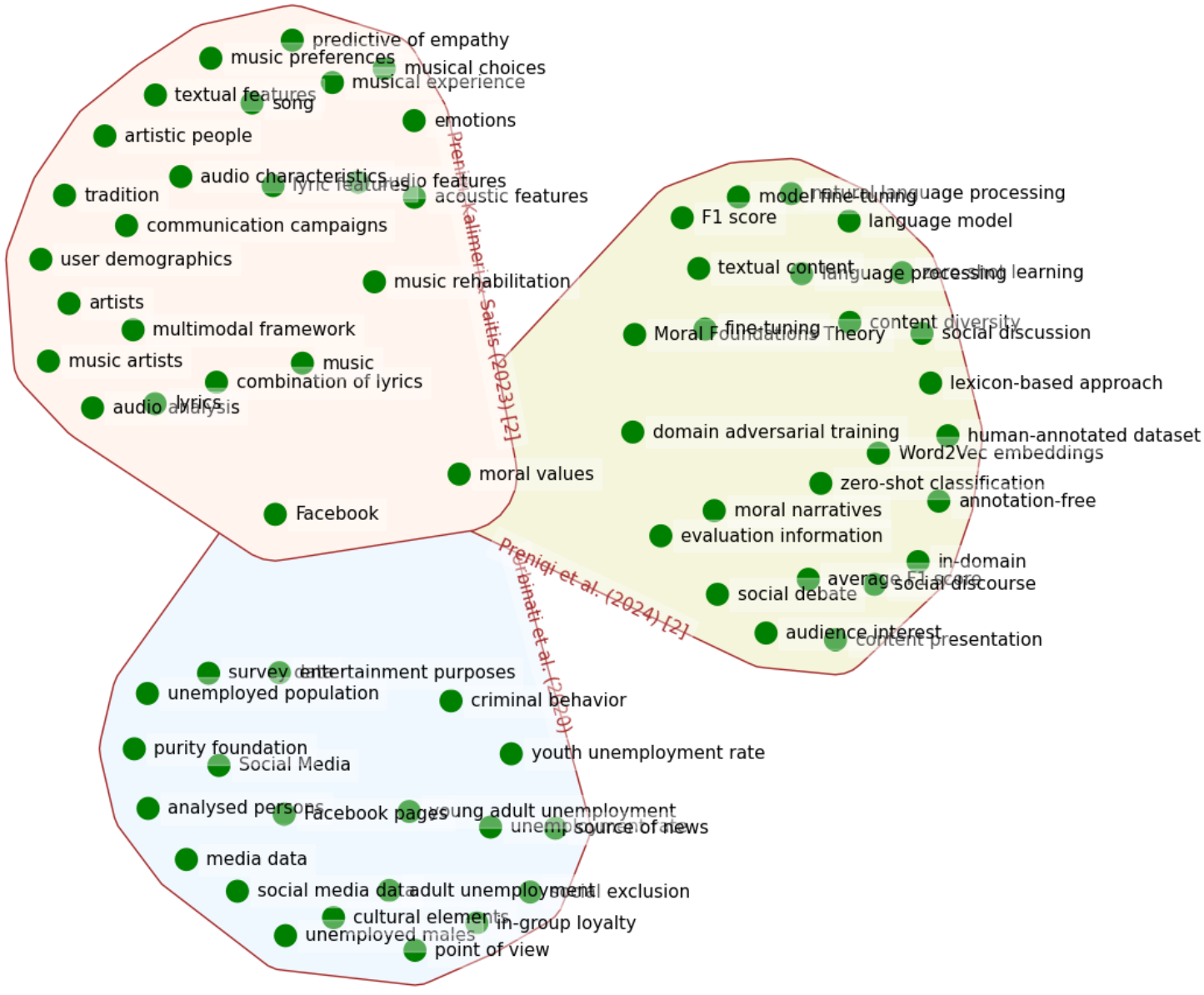

Jaccard Index Heatmap of Concept Hyperedge Intersections
Zitnik et al. (2024)
Zitnik et al. (2018) [2]
Johnson et al. (2024)
Huang et al. (2024) [2]
Gonzalez et al. (2024)
Lin et al. (2023)
Johnson et al. (2023)
Huang et al. (2020) [4]
Huang et al. (2020) [2]
Hyperedges
Hyperedges
Jaccard Index
1.0
0.8
0.6
0.4
0.2
0.0
1.000 0.043 0.022 0.026 0.000 0.023 0.022 0.040 0.038
0.043 1.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
0.022 0.000 1.000 0.158 0.044 0.089 0.625 0.056 0.053
0.026 0.000 0.158 1.000 0.054 0.108 0.158 0.065 0.061
0.000 0.000 0.044 0.054 1.000 0.000 0.044 0.040 0.038
0.023 0.000 0.089 0.108 0.000 1.000 0.065 0.019 0.018
0.022 0.000 0.625 0.158 0.044 0.065 1.000 0.056 0.053
0.040 0.000 0.056 0.065 0.040 0.019 0.056 1.000 0.806
0.038 0.000 0.053 0.061 0.038 0.018 0.053 0.806 1.000

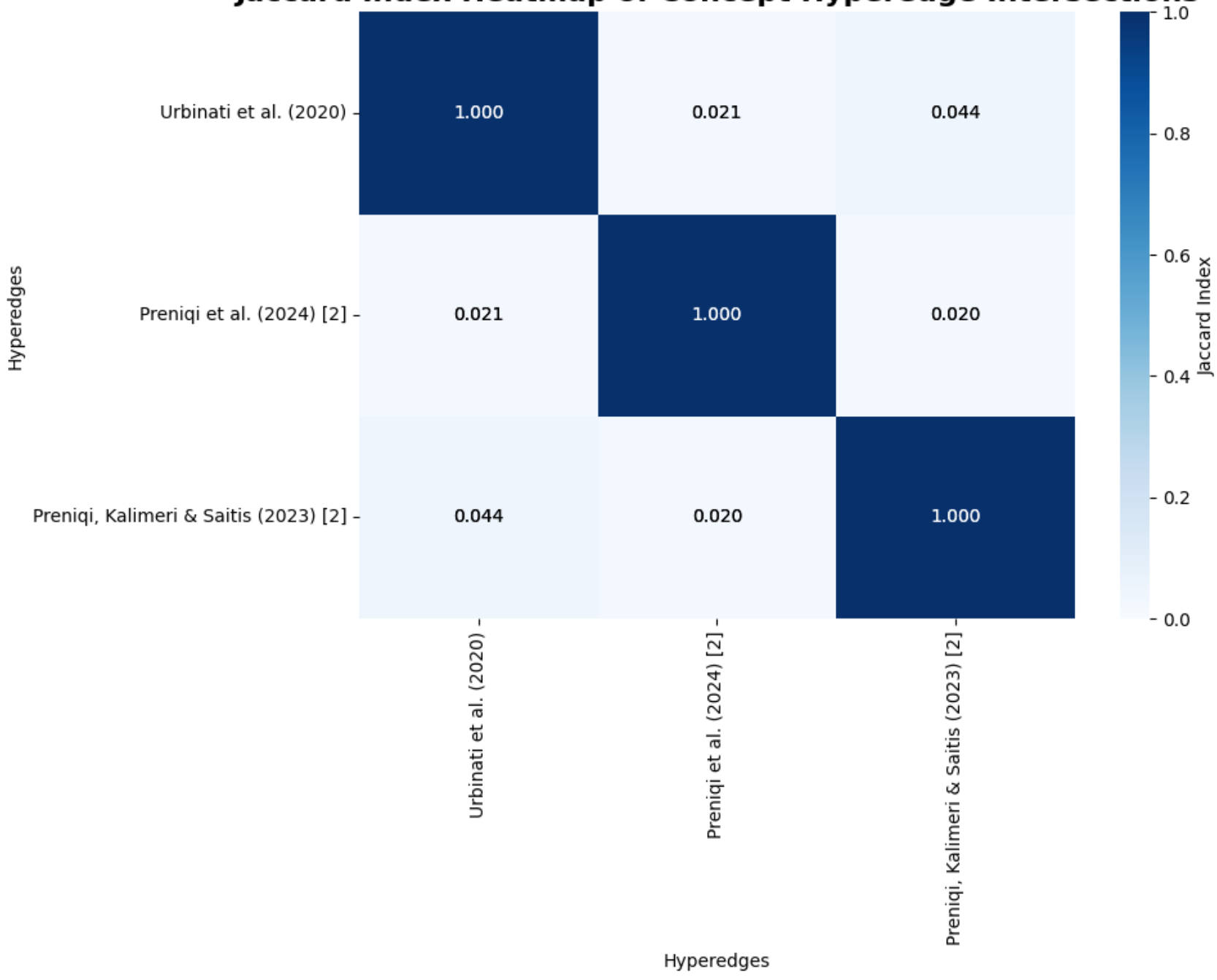
The Hypergraph of Research Areas of Marinka Zitnik in 8 Top Citations



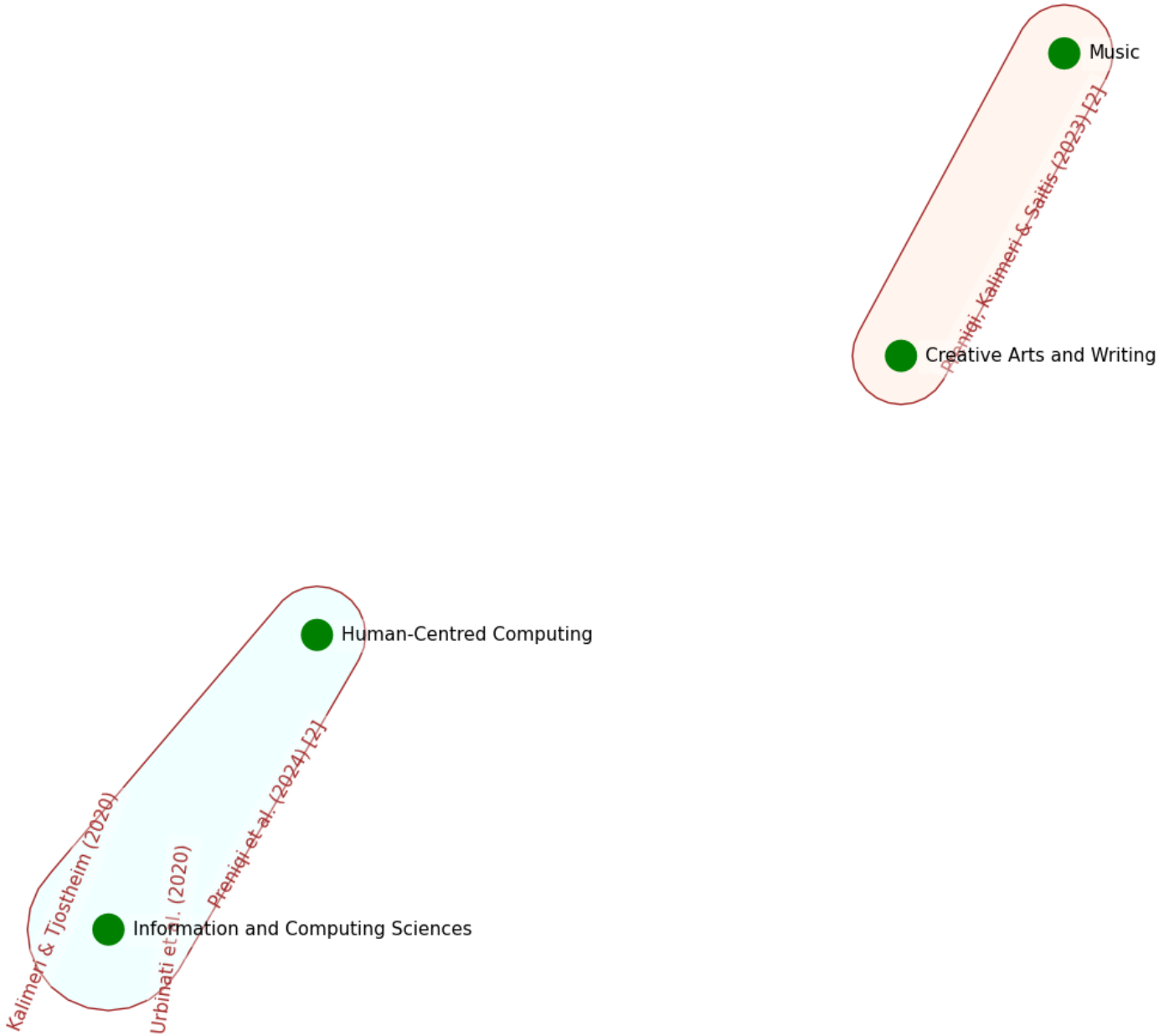

Jaccard Index Heatmap of Research Area Hyperedge Intersections
Zitnik et al. (2018) [2]
Huang et al. (2020) [2]
Huang et al. (2020) [4]
Lin et al. (2023)
Johnson et al. (2023)
Gonzalez et al. (2024)
Zitnik et al. (2024)
Johnson et al. (2024)
Hyperedges
1.000 0.100 0.111 0.750 0.200 0.143 0.167 0.000
0.100 1.000 0.857 0.111 0.286 0.571 0.250 0.286
0.111 0.857 1.000 0.125 0.143 0.667 0.286 0.333
0.750 0.111 0.125 1.000 0.250 0.167 0.200 0.000
0.200 0.286 0.143 0.250 1.000 0.200 0.250 0.000
0.143 0.571 0.667 0.167 0.200 1.000 0.400 0.200
0.167 0.250 0.286 0.200 0.250 0.400 1.000 0.250
0.000 0.286 0.333 0.000 0.000 0.200 0.250 1.000
Zitnik et al. (2018) [2]
Huang et al. (2020) [2]
Huang et al. (2020) [4]
Lin et al. (2023)
Johnson et al. (2023)
Gonzalez et al. (2024)
Zitnik et al. (2024)
Johnson et al. (2024)
Hyperedges
Jaccard Index
1.0
0.8
0.6
0.4
0.2
0.0

The Hypergraph of Concepts of Kyriaki Kalimeri in 3 Top Citations




Jaccard Index Heatmap of Concept Hyperedge Intersections
Urbinati et al. (2020)
Preniqi et al. (2024) [2]
Preniqi, Kalimeri & Saitis (2023) [2]
1.000
0.021
0.044
0.021
1.000
0.020
0.044
0.020
1.000
Hyperedges
Hyperedges
Urbinati et al. (2020)
Preniqi et al. (2024) [2]
Preniqi, Kalimeri & Saitis (2023) [2]
Jaccard Index
1.0
0.8
0.6
0.4
0.2
0.0

The Hypergraph of Research Areas of Kyriaki Kalimeri in 4 Top Citations




Jaccard Index Heatmap of Research Area Hyperedge Intersections
Hyperedges
Kalimeri & Tjostheim (2020)
Urbinati et al. (2020)
Preniqi, Kalimeri & Saitis (2023) [2]
Preniqi et al. (2024) [2]
1.000
1.000
0.000
0.500
1.000
1.000
0.000
0.500
0.000
0.000
1.000
0.000
0.500
0.500
0.000
1.000
Kalimeri & Tjostheim (2020)
Urbinati et al. (2020)
Preniqi, Kalimeri & Saitis (2023) [2]
Preniqi et al. (2024) [2]
Hyperedges
Jaccard Index
1.0
0.8
0.6
0.4
0.2
0.0