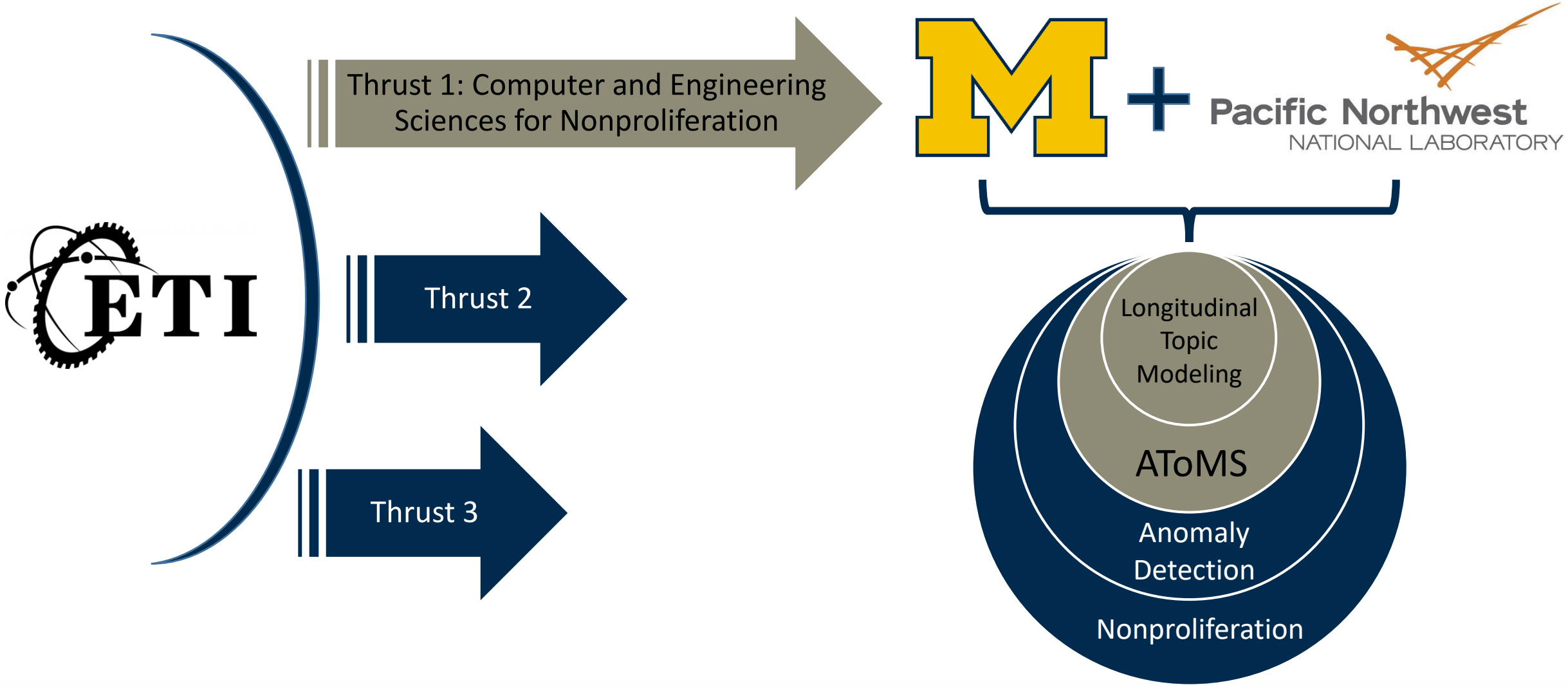
AToMS: Author-Topic Manifold Summarization for Interpretable Author-Collaboration Forecasting

Conrad D. Hougen[§], Karl T. Pazdernik[⁑], Alfred O. Hero[§]

University of Michigan[§], Pacific Northwest National Laboratory[⁑]

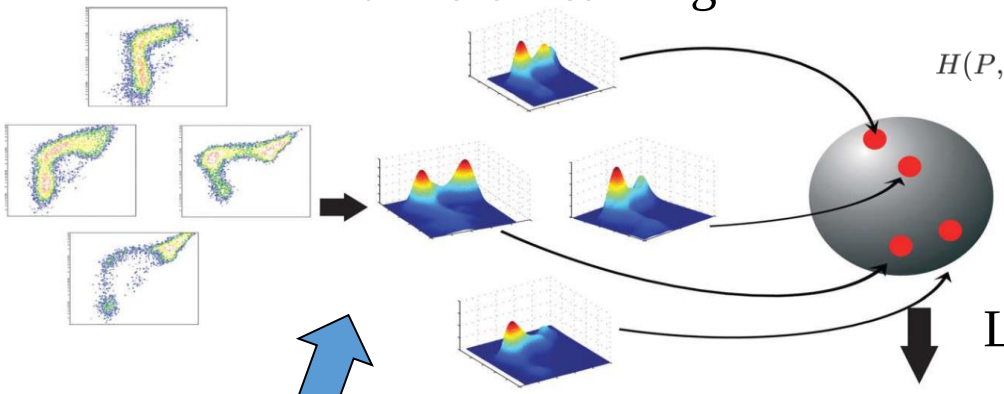
Date 02/08/23

PNNL-SA-181770


Thrust 1: Computer and Engineering Sciences for Nonproliferation
Thrust 2
Thrust 3
Pacific Northwest
NATIONAL LABORATORY
Longitudinal Topic Modeling
AToMS
Anomaly Detection
Nonproliferation

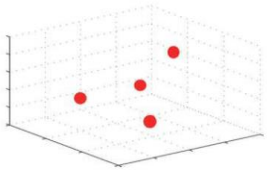
Foundations

Manifold Learning

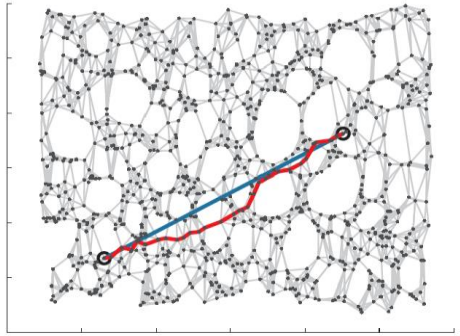


$$H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{n=1}^{N}(\sqrt{p_n}-\sqrt{q_n})^2}$$

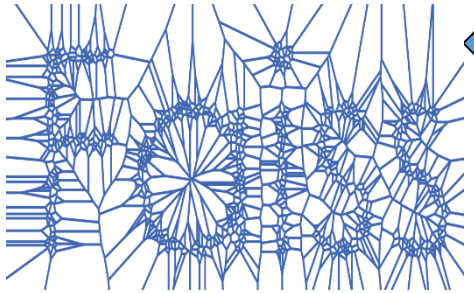
Low-Dimensional Embedding

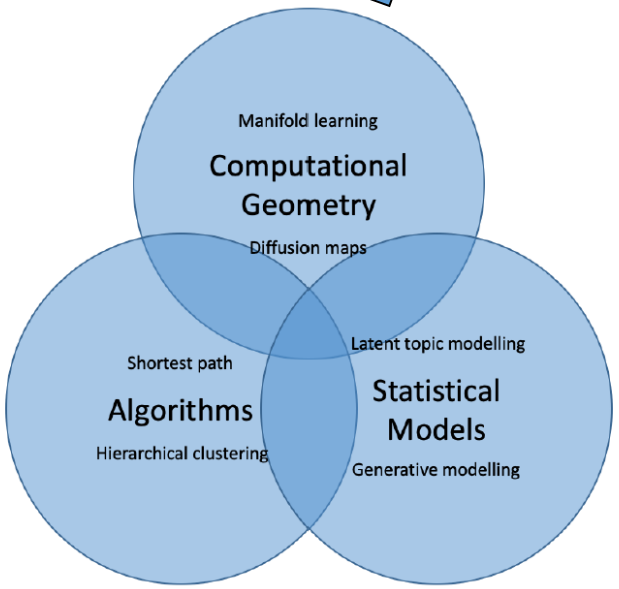


Dijkstra's Algorithm

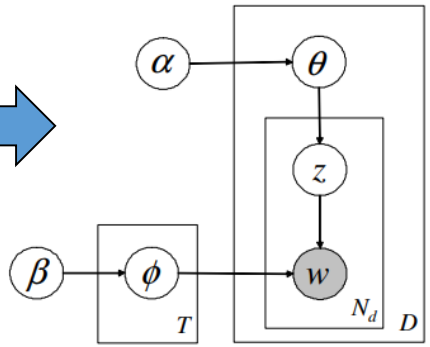


Similarity Search

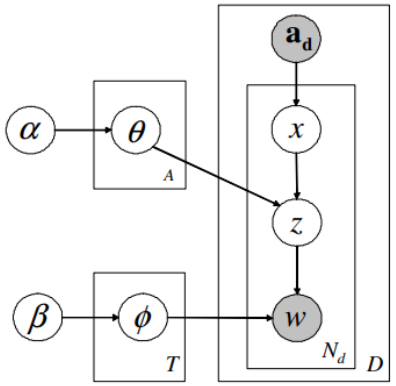




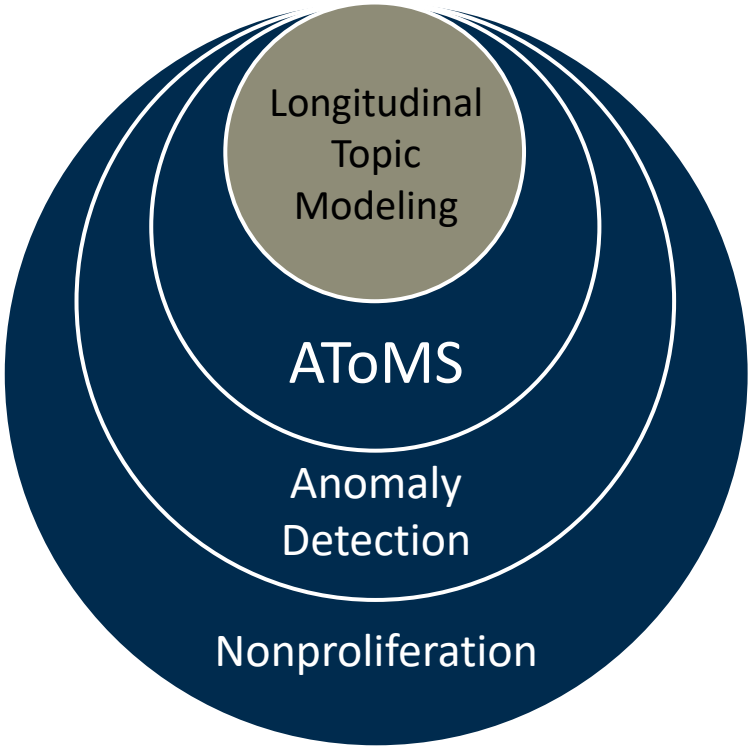
LDA



Author-Topic Model



GDLTM: A Geometry-Driven Longitudinal Topic Model



[1] Wang, Y., Hougen, C., Oselio, B., Dempsey, W., & Hero, A. "A Geometry-Driven Longitudinal Topic Model." Harvard Data Science Review (2021).

GDLTM: A Geometry-Driven Longitudinal Topic Model

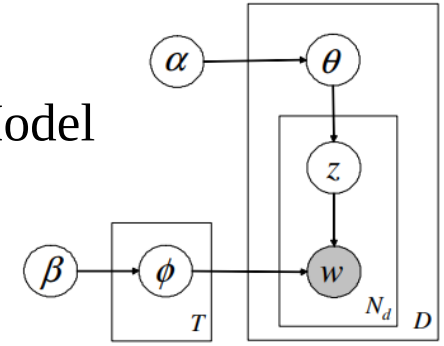
Temporal Smoothing

d_1	d_2	d_3	d_4	d_5
Doc1	Doc6	Doc11	Doc16	Doc21
Doc2	Doc7	Doc12	Doc17	Doc22
Doc3	Doc8	Doc13	Doc18	Doc23
Doc4	Doc9	Doc14	Doc19	Doc24
Doc5	Doc10	Doc15	Doc20	Doc25

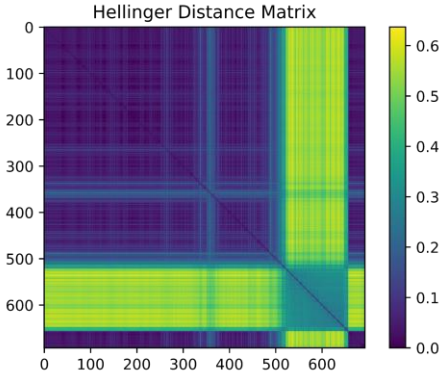
r_1	r_2	r_3	r_4	r_5
Doc1	Doc6	Doc11	Doc17	Doc23
Doc2	Doc8	Doc12	Doc19	
Doc3	Doc9	Doc13		
Doc4	Doc10			
Doc5				

s_1	s_2	s_3	s_4	s_5
Doc1	Doc6	Doc11	Doc17	Doc23
Doc2	**Doc7**	Doc12	Doc19	**Doc24**
~~Doc3~~	Doc8	Doc13	**Doc20**	
Doc4	Doc9	**Doc15**		
Doc5	Doc10			

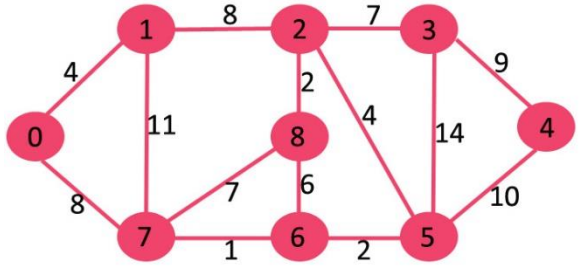
Topic Model



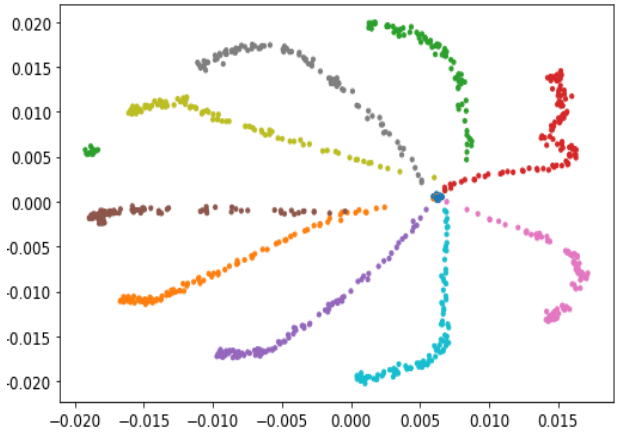
Distance Matrix



Shortest Path Algorithm



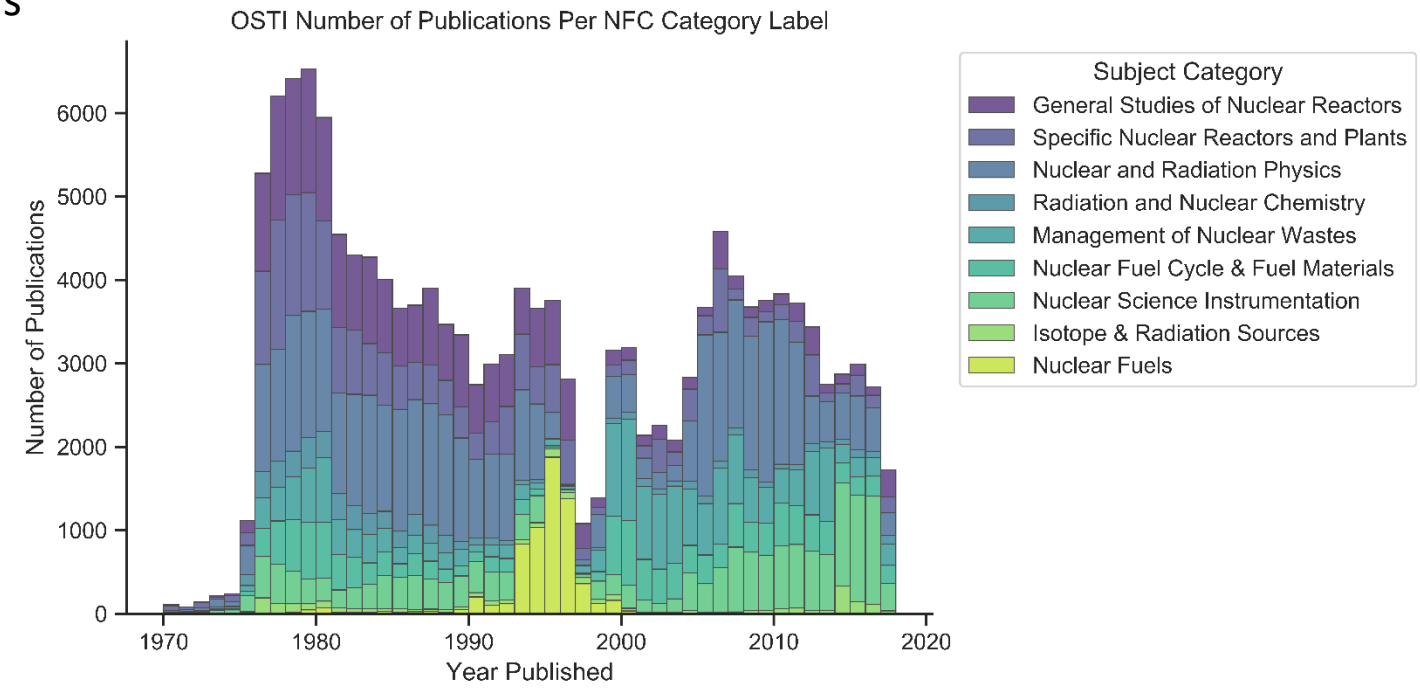
Low-Dimensional Embedding



[1] Wang, Y., Hougen, C., Oselio, B., Dempsey, W., & Hero, A. "A Geometry-Driven Longitudinal Topic Model." Harvard Data Science Review (2021).

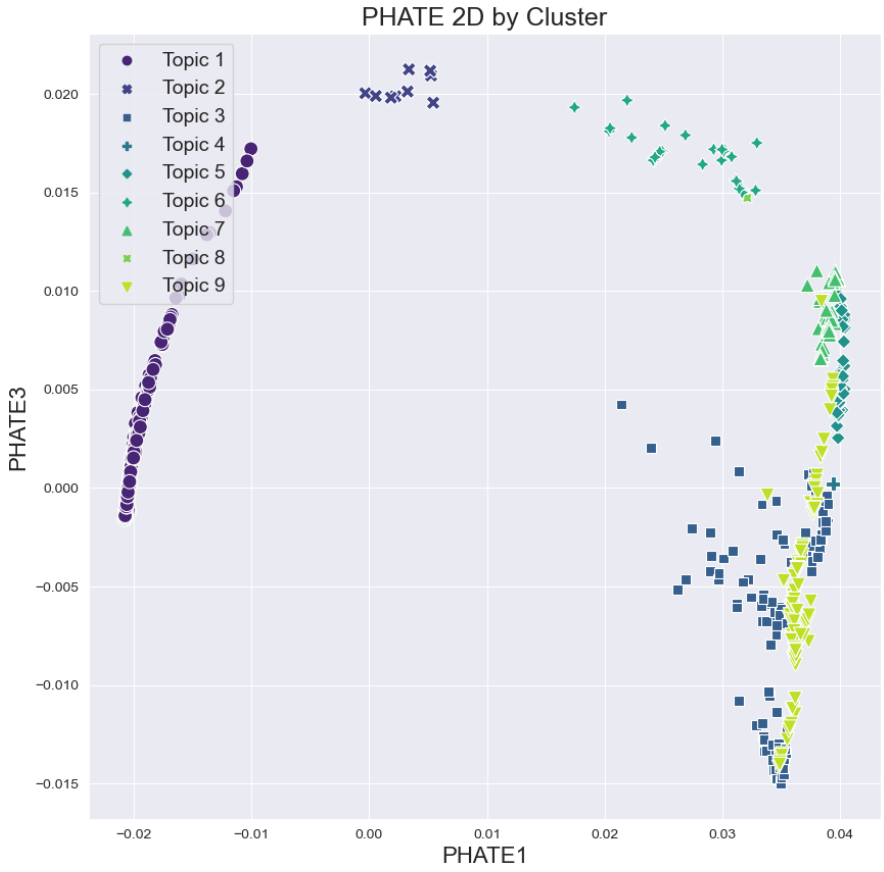
GDLTM: A Geometry-Driven Longitudinal Topic Model

- GDLTM on OSTI dataset of manuscript abstracts
- Filtered on **Nuclear Fuel Cycle** labeled papers
  - ~150,000 manuscripts
  - ~360,000 authors

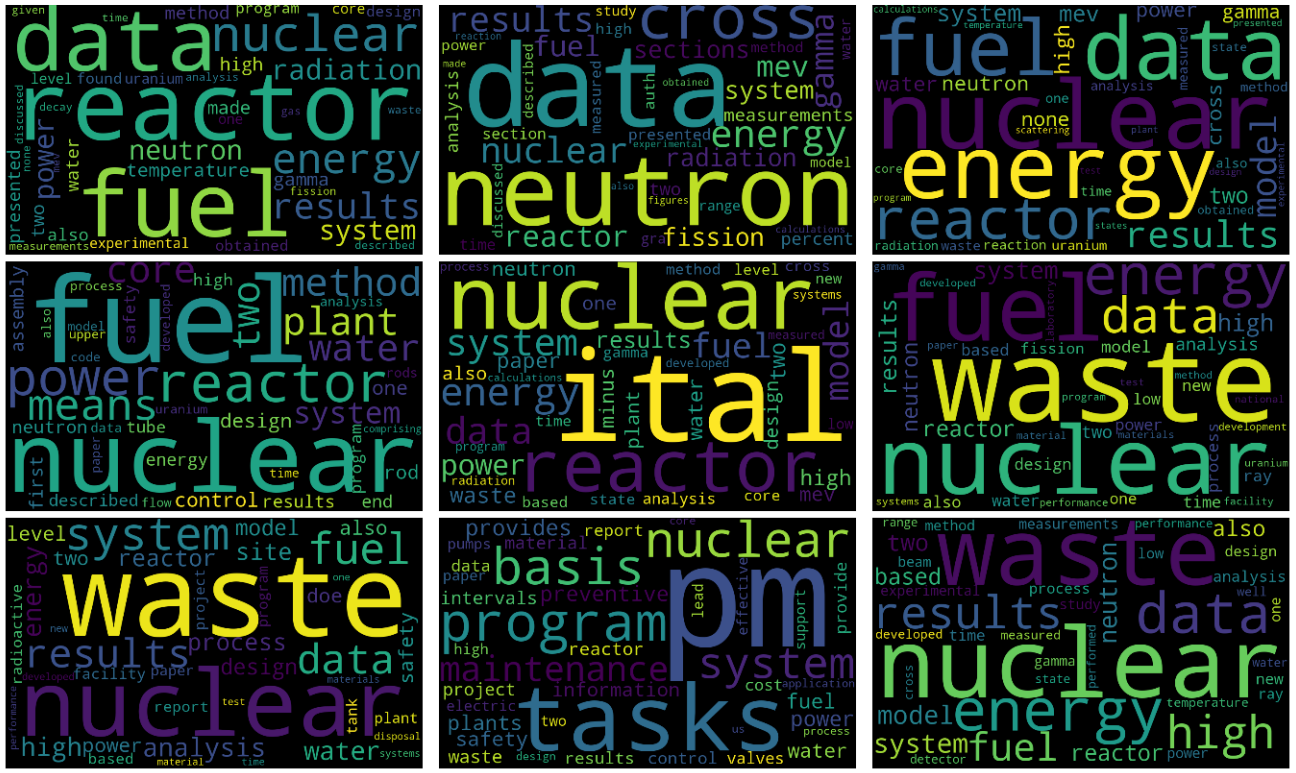


[1] Wang, Y., Hougen, C., Oselio, B., Dempsey, W., & Hero, A. "A Geometry-Driven Longitudinal Topic Model." Harvard Data Science Review (2021).

GDLTM: A Geometry-Driven Longitudinal Topic Model

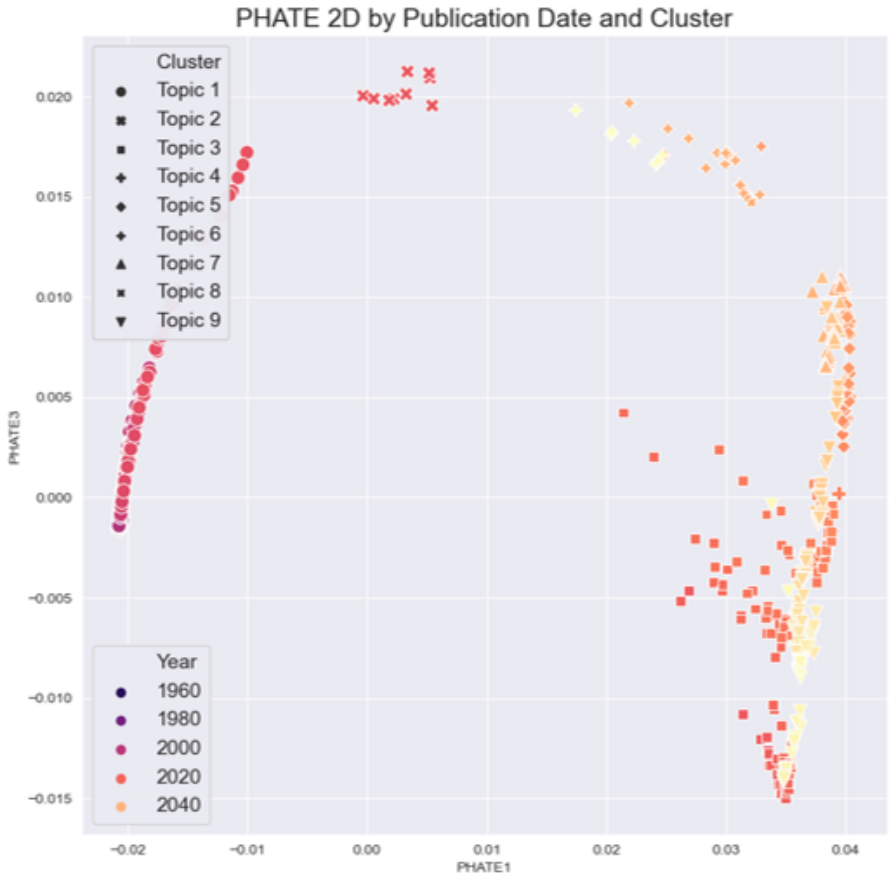


Clustered Topic Wordclouds

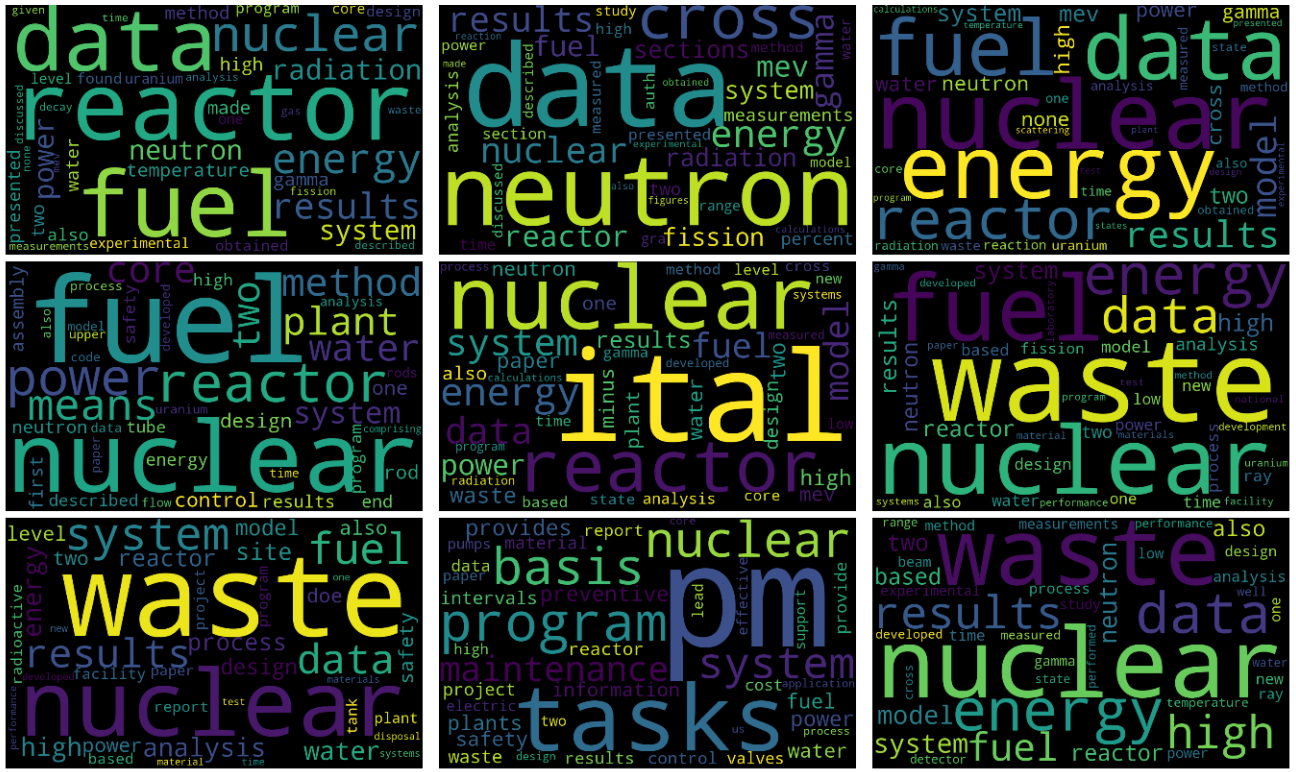


[1] Wang, Y., Hougen, C., Oselio, B., Dempsey, W., & Hero, A. "A Geometry-Driven Longitudinal Topic Model." Harvard Data Science Review (2021).

GDLTM: A Geometry-Driven Longitudinal Topic Model



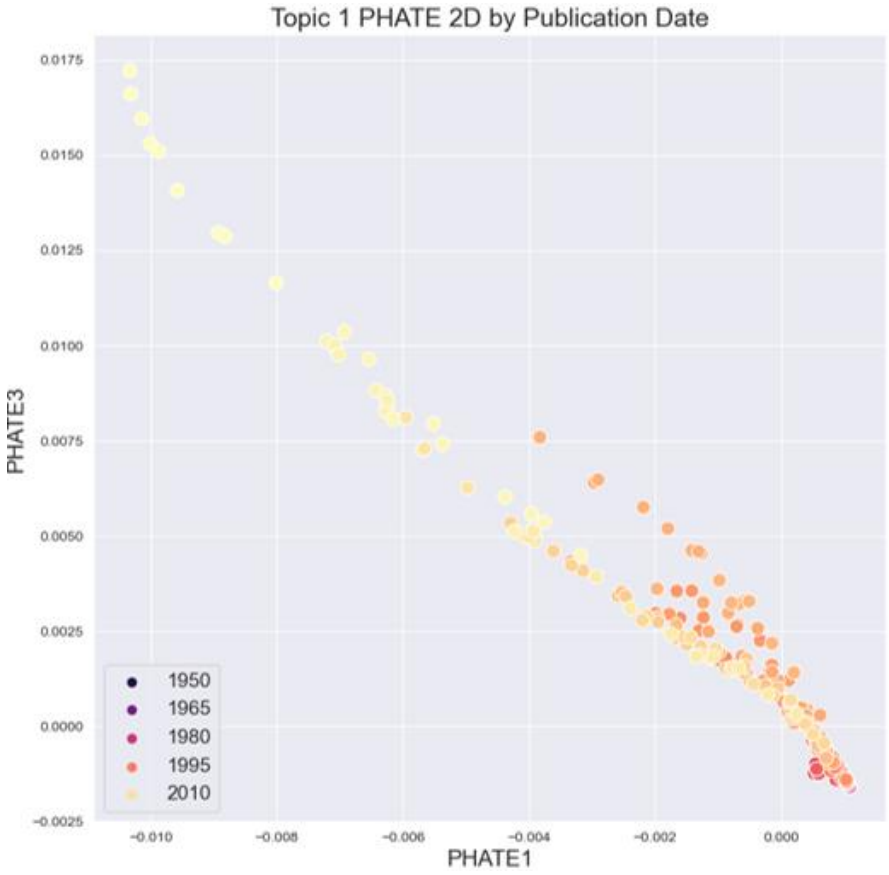
Clustered Topic Wordclouds



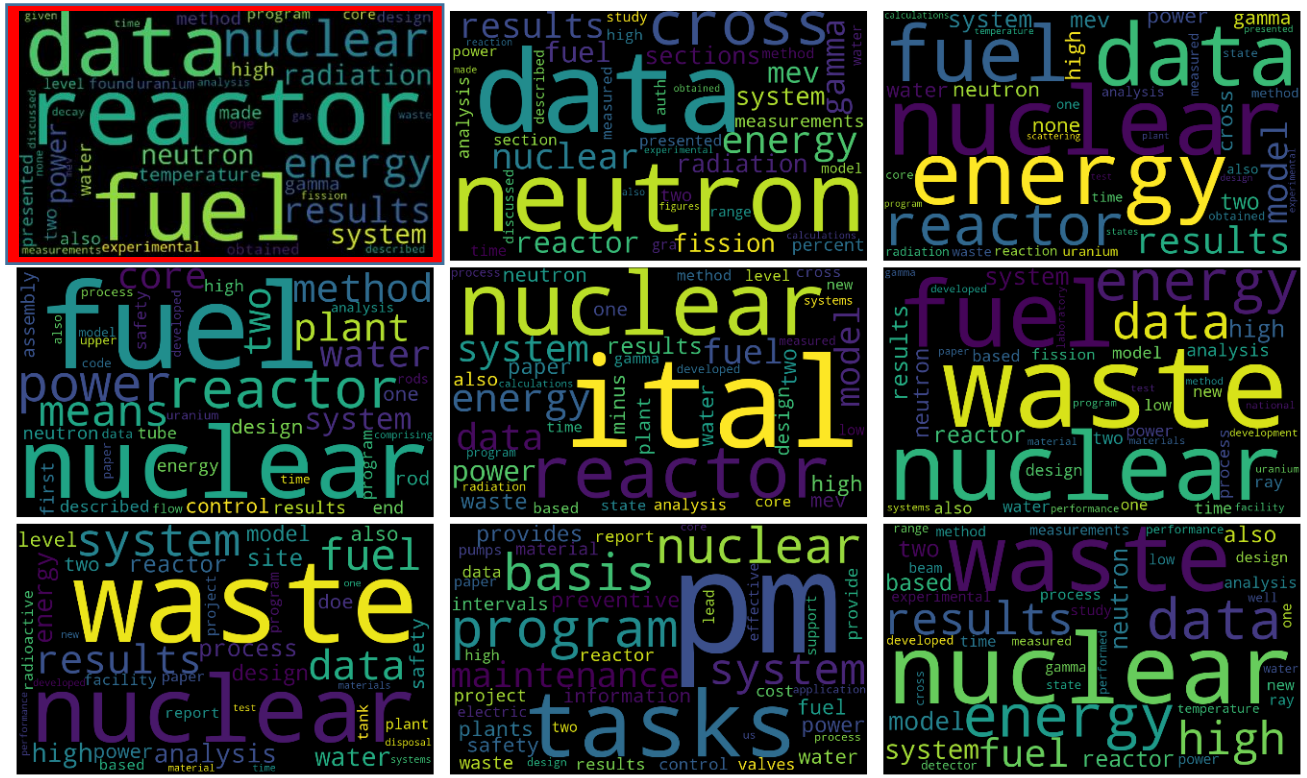
[1] Wang, Y., Hougen, C., Oselio, B., Dempsey, W., & Hero, A. "A Geometry-Driven Longitudinal Topic Model." Harvard Data Science Review (2021).

PNNL-SA-181770

GDLTM: A Geometry-Driven Longitudinal Topic Model

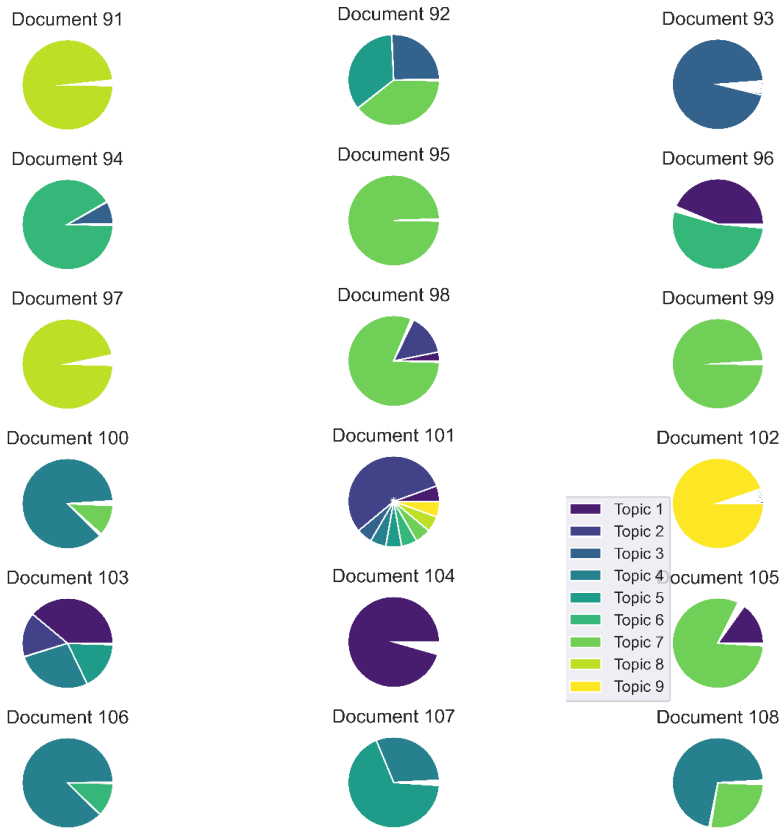


Clustered Topic Wordclouds

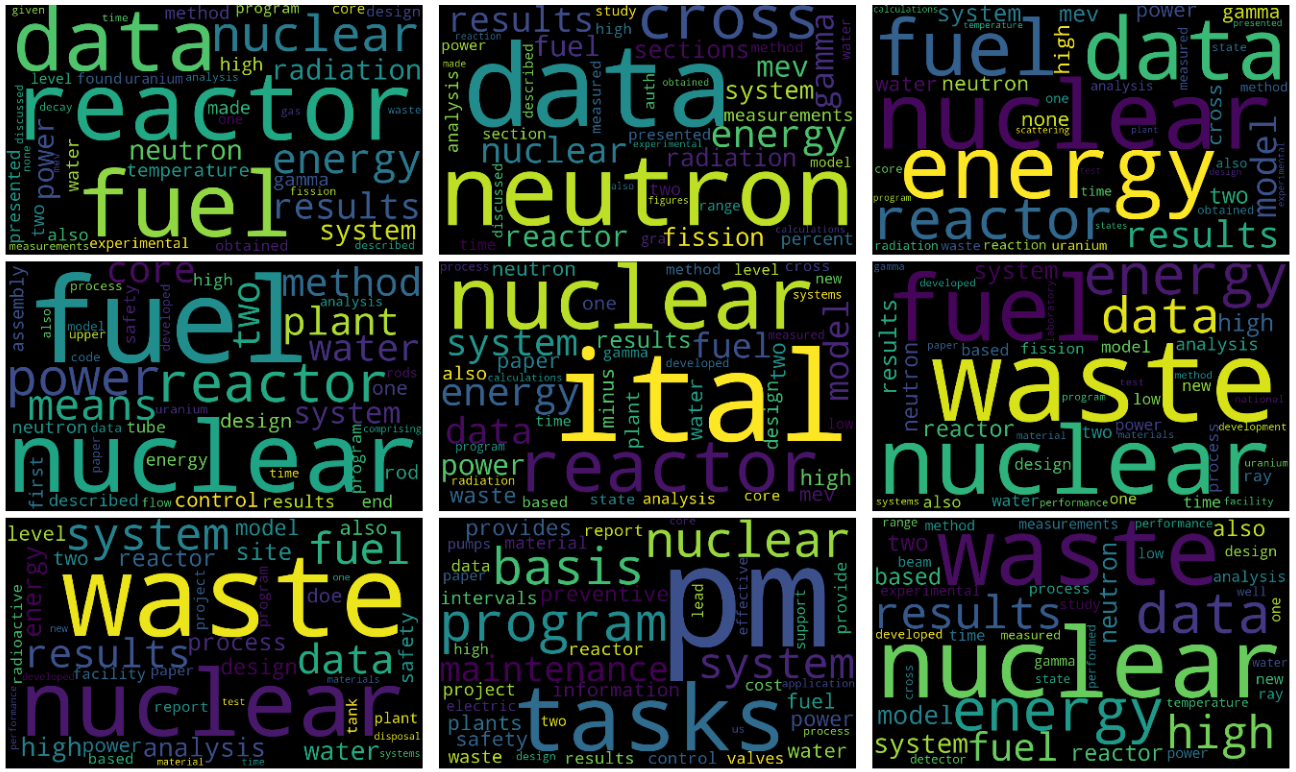


[1] Wang, Y., Hougen, C., Oselio, B., Dempsey, W., & Hero, A. "A Geometry-Driven Longitudinal Topic Model." Harvard Data Science Review (2021).

GDLTM: A Geometry-Driven Longitudinal Topic Model



Clustered Topic Wordclouds

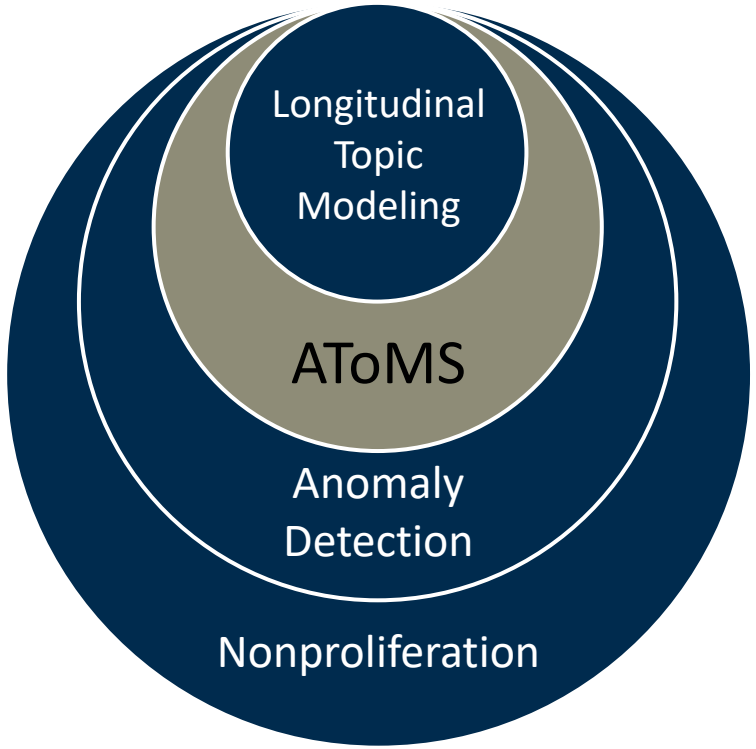


[1] Wang, Y., Hougen, C., Oselio, B., Dempsey, W., & Hero, A. "A Geometry-Driven Longitudinal Topic Model." Harvard Data Science Review (2021).

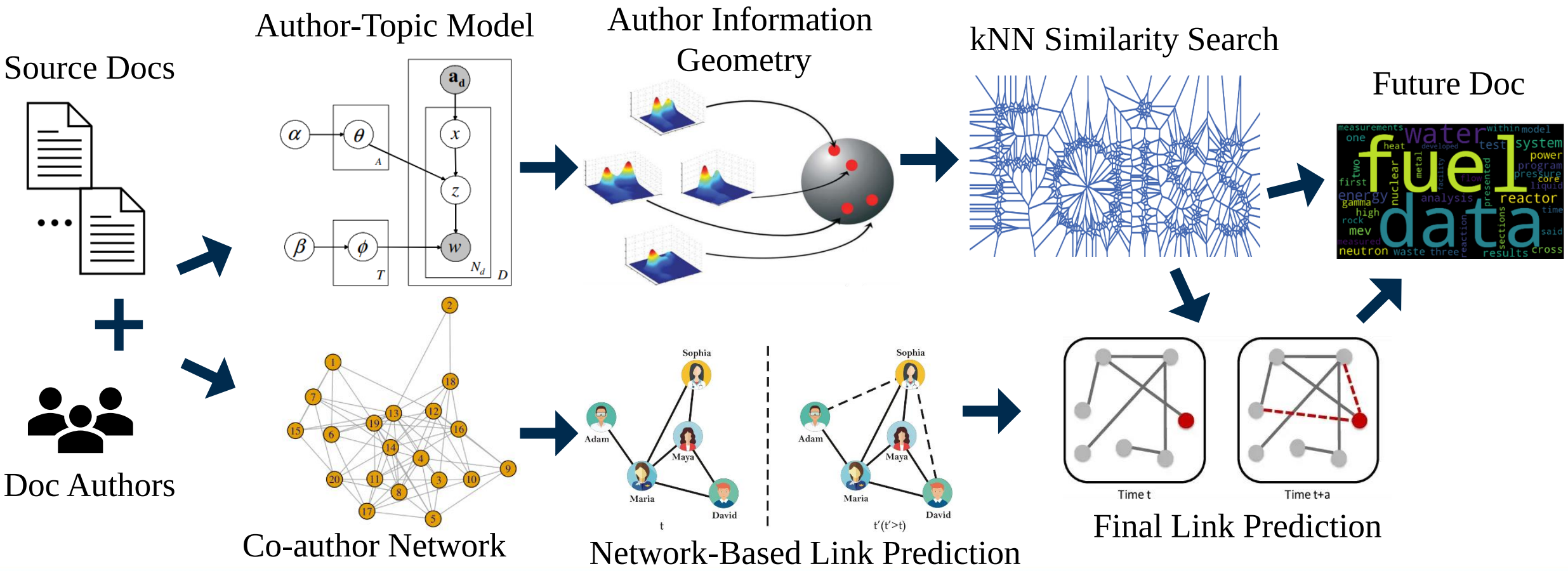
National Nuclear Security Administration

PNNL-SA-181770

AToMS: Author Topic Manifold Summarization



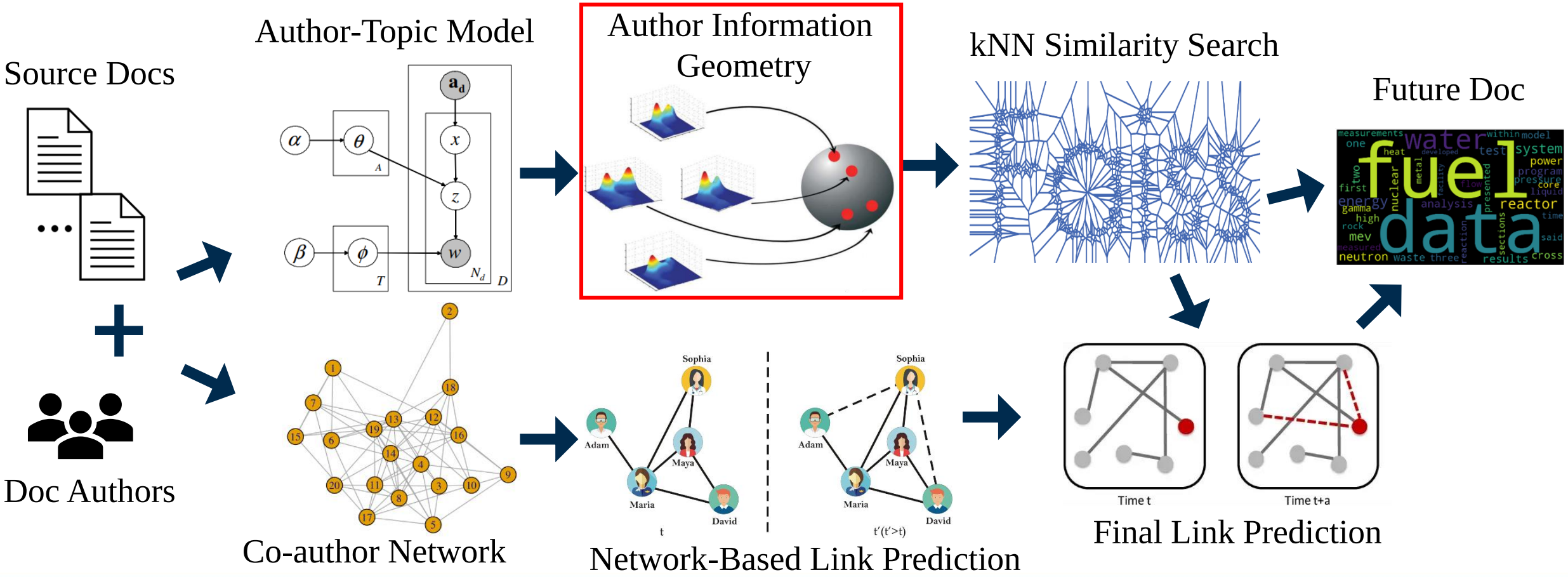
ATOMS: Author Topic Manifold Summarization



[2] Ahmad, I., Akhtar, M.U., Noor, S. *et al.* Missing Link Prediction using Common Neighbor and Centrality based Parameterized Algorithm. *Sci Rep* 10, 364 (2020).

[3] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In Proceedings of the 20th conference on Uncertainty in artificial intelligence (UAI '04). AUAI Press, Arlington, Virginia, USA, 487–494.

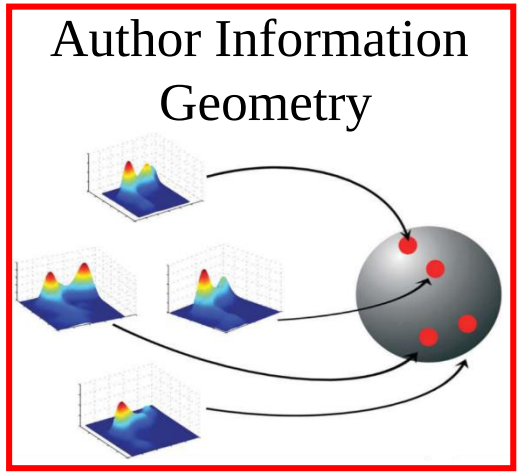
ATOMS: Author Topic Manifold Summarization



[2] Ahmad, I., Akhtar, M.U., Noor, S. *et al.* Missing Link Prediction using Common Neighbor and Centrality based Parameterized Algorithm. *Sci Rep* 10, 364 (2020).

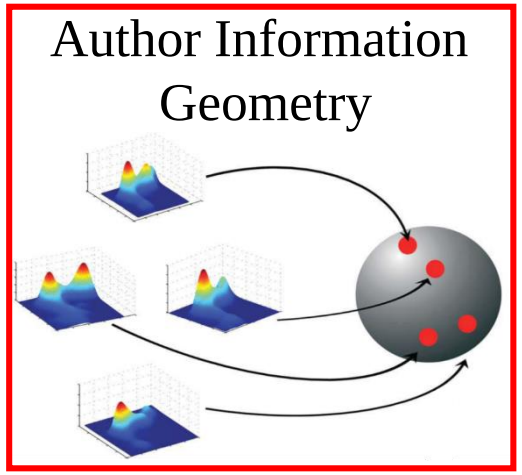
[3] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In Proceedings of the 20th conference on Uncertainty in artificial intelligence (UAI '04). AUAI Press, Arlington, Virginia, USA, 487–494.

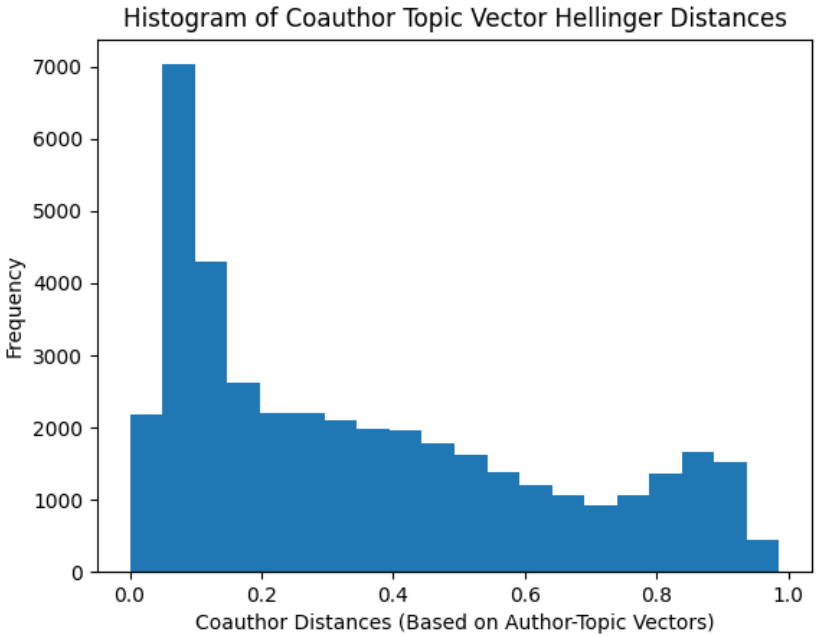
ATOMS: Author Topic Manifold Summarization

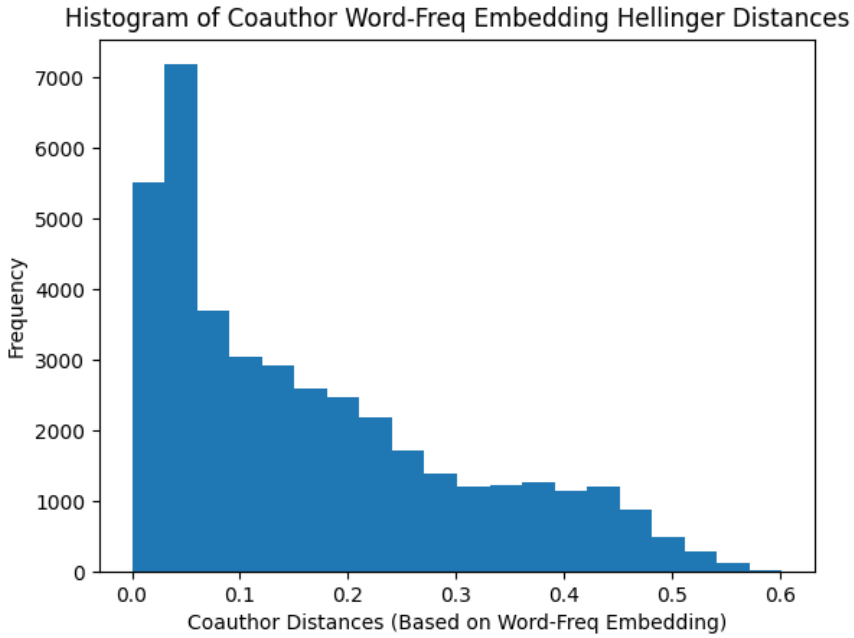


- Important Question: How should we represent authors?
  - Representation depends on distance metric
  - Hellinger distance: $H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{n=1}^{N}\left(\sqrt{p_n} - \sqrt{q_n}\right)^2}$
  - Topic vectors are not always well-separated
  - Alternative symmetric distance metrics may have disadvantages
    - E.g. Wasserstein, cosine, etc.
    - Depends on the author embedding

AToMS Author Vectors

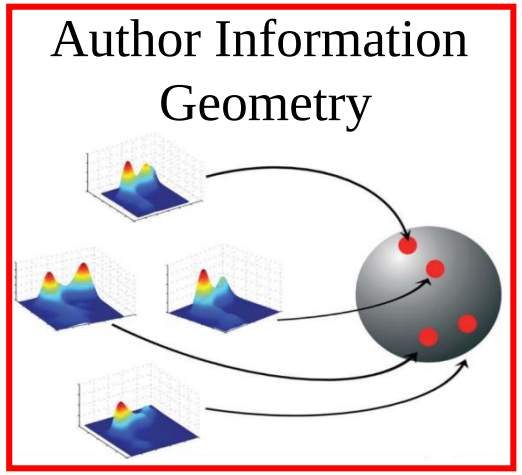


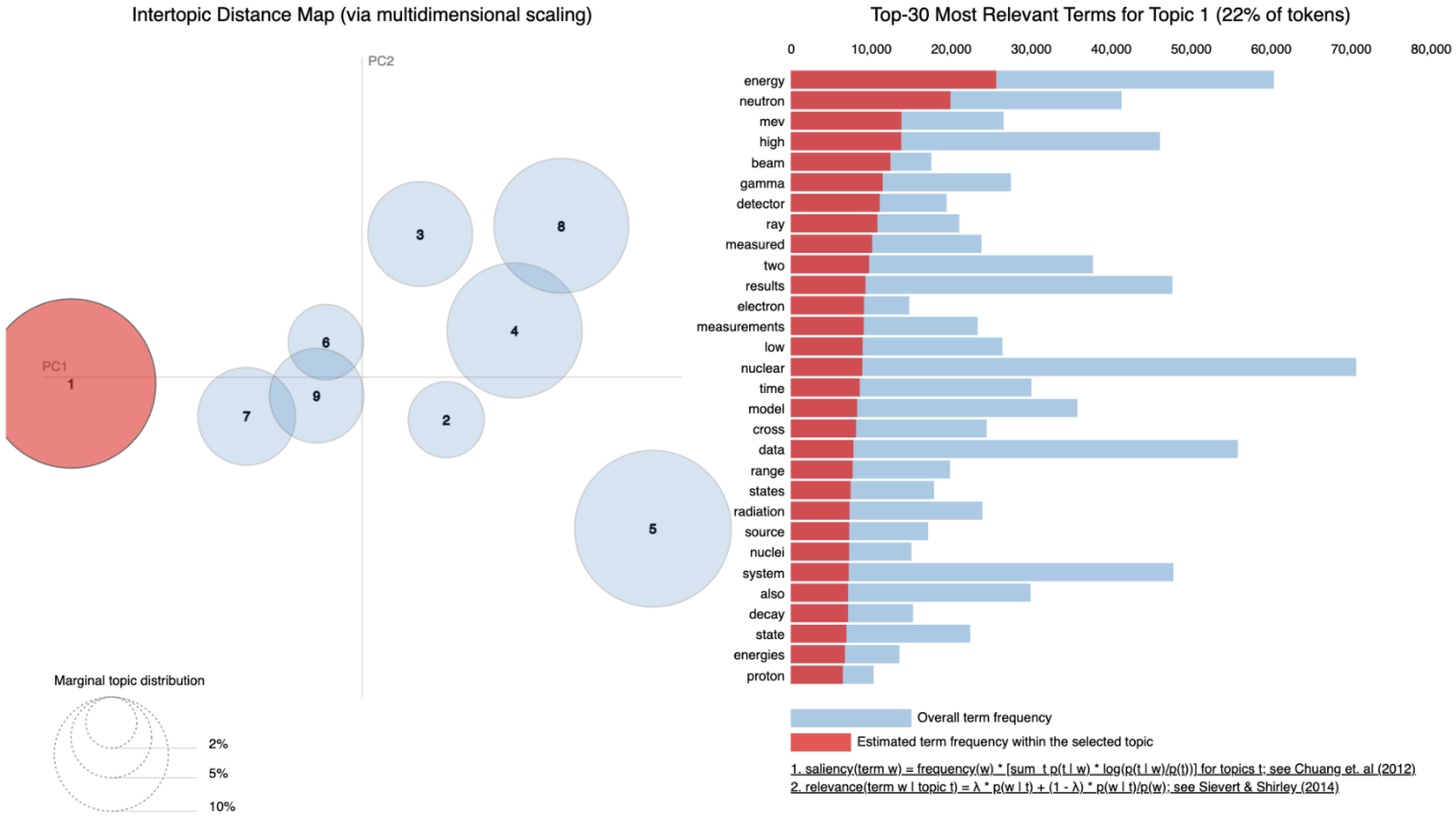




- Topic vectors are not always well-separated

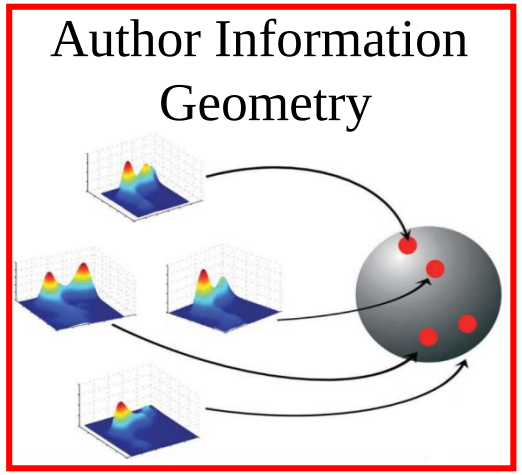
AToMS: Author Topic Manifold Summarization

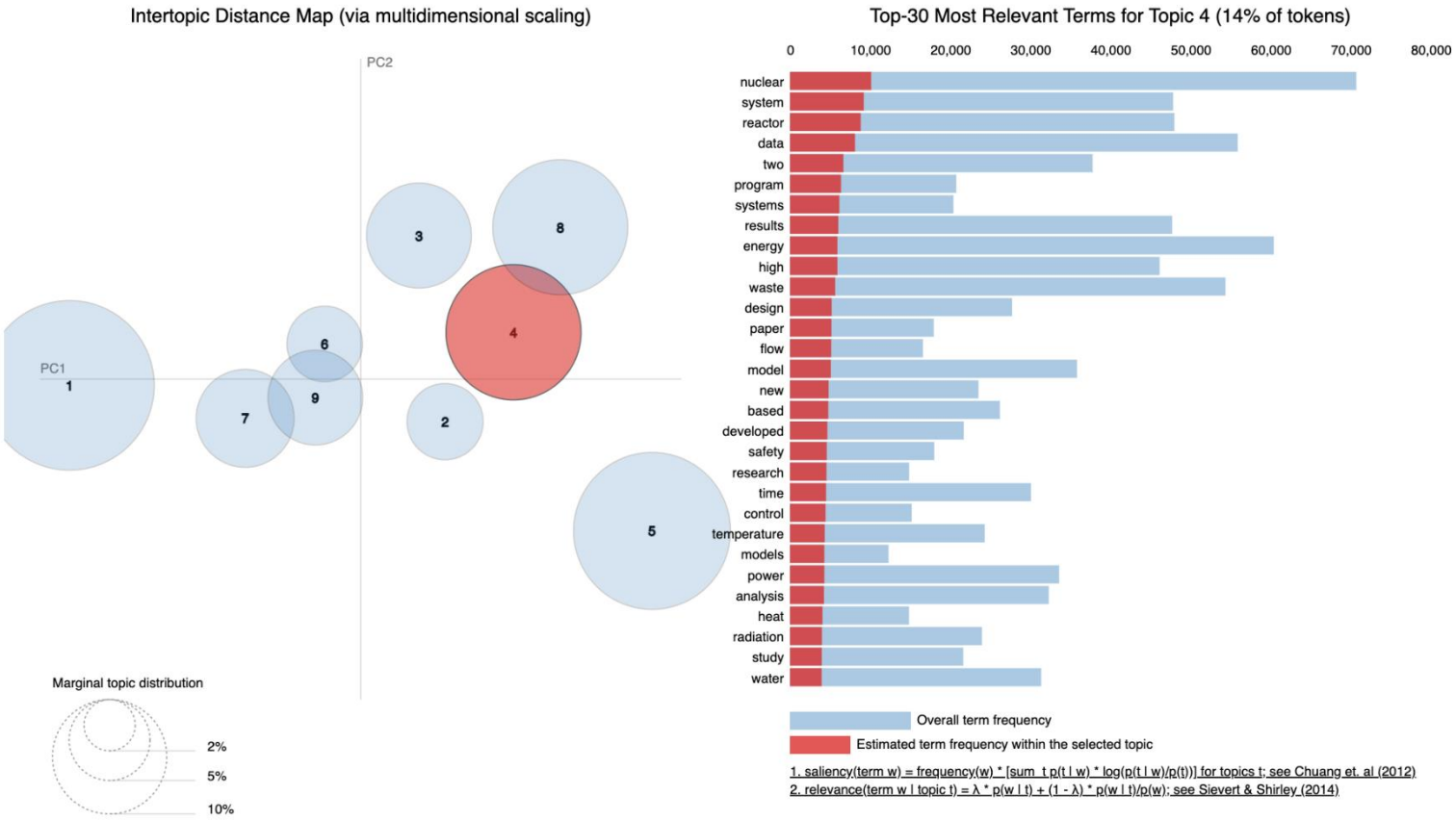




- Topic vectors are not always well-separated

AToMS: Author Topic Manifold Summarization

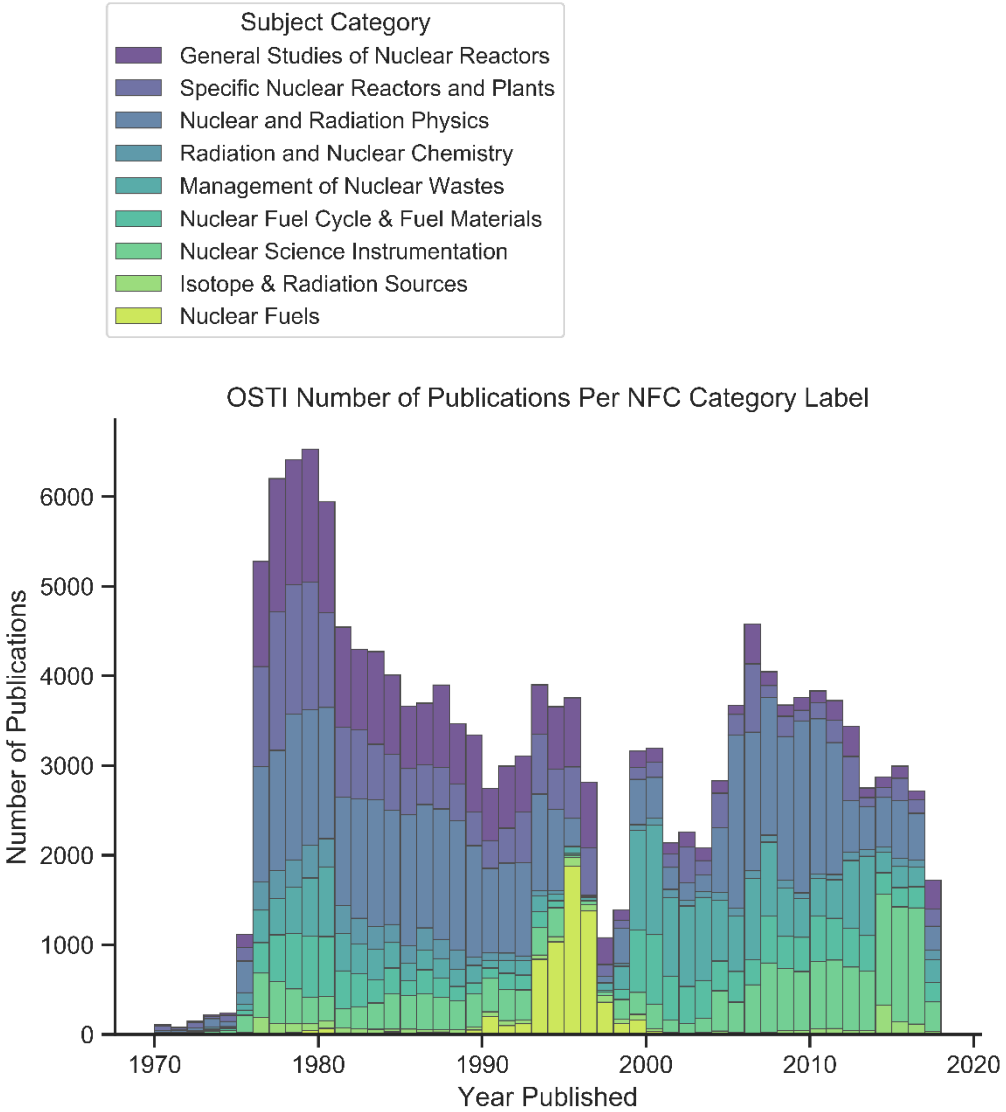




- Topic vectors are not always well-separated

Experimental Setup

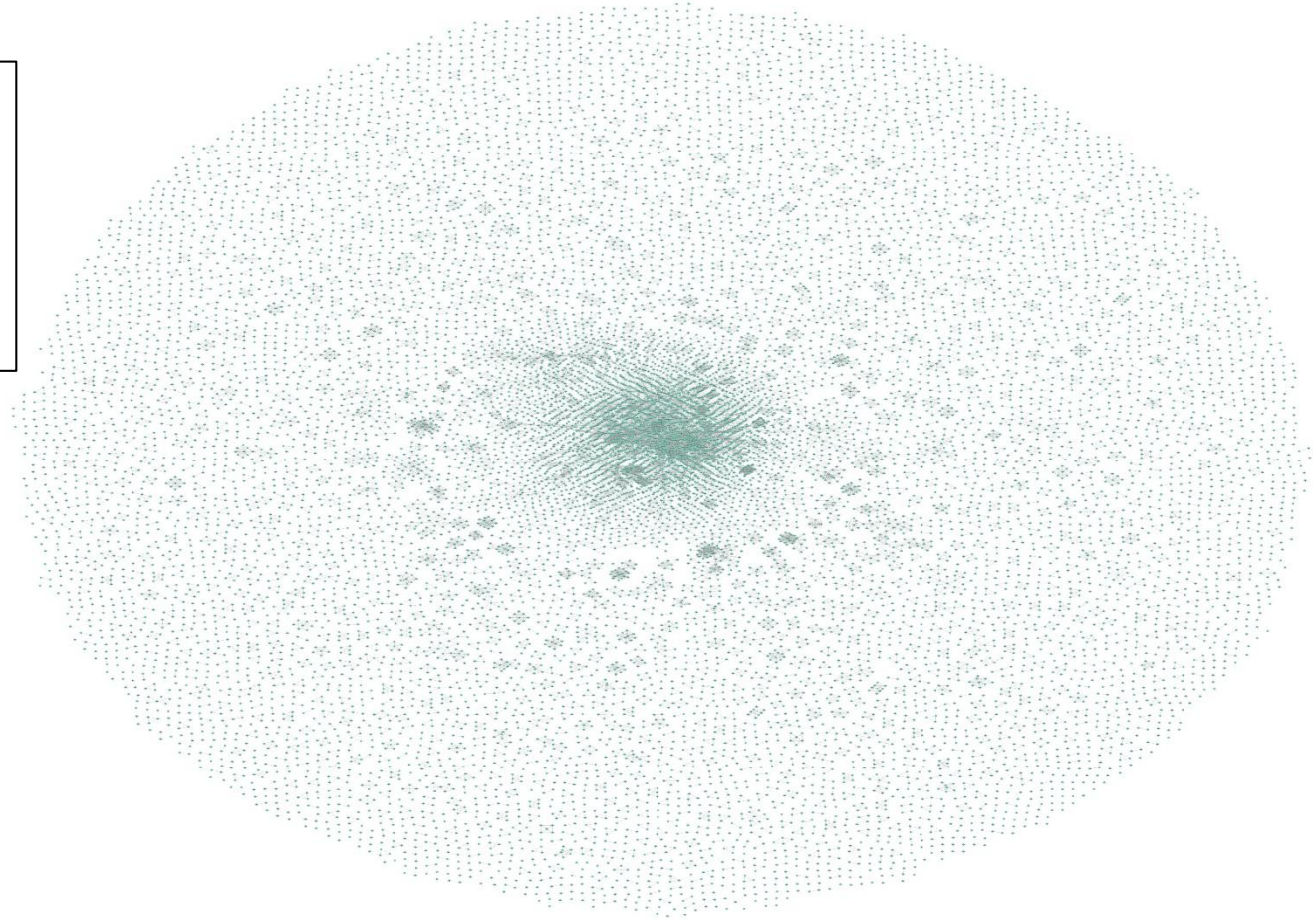
- OSTI (DOE Office of Scientific and Technical Information)
  - Database of articles from DOE-funded R&D
  - Title, author, OSTI subject label, abstract text, etc.
- Nuclear Fuel Cycle Documents
  - Filtered from OSTI using subject labels
  - 9 OSTI subject labels are considered NFC related
- Preprocessing
  - Filter out “stop words” – standard practice in NLP
  - Author entity disambiguation
  - Remove documents with missing abstract or authors
- Experimental Setup
  - Reduce corpus to specific subsets for train and test
  - Consider only authors that exist in the training set
  - Predict links between pairs of authors

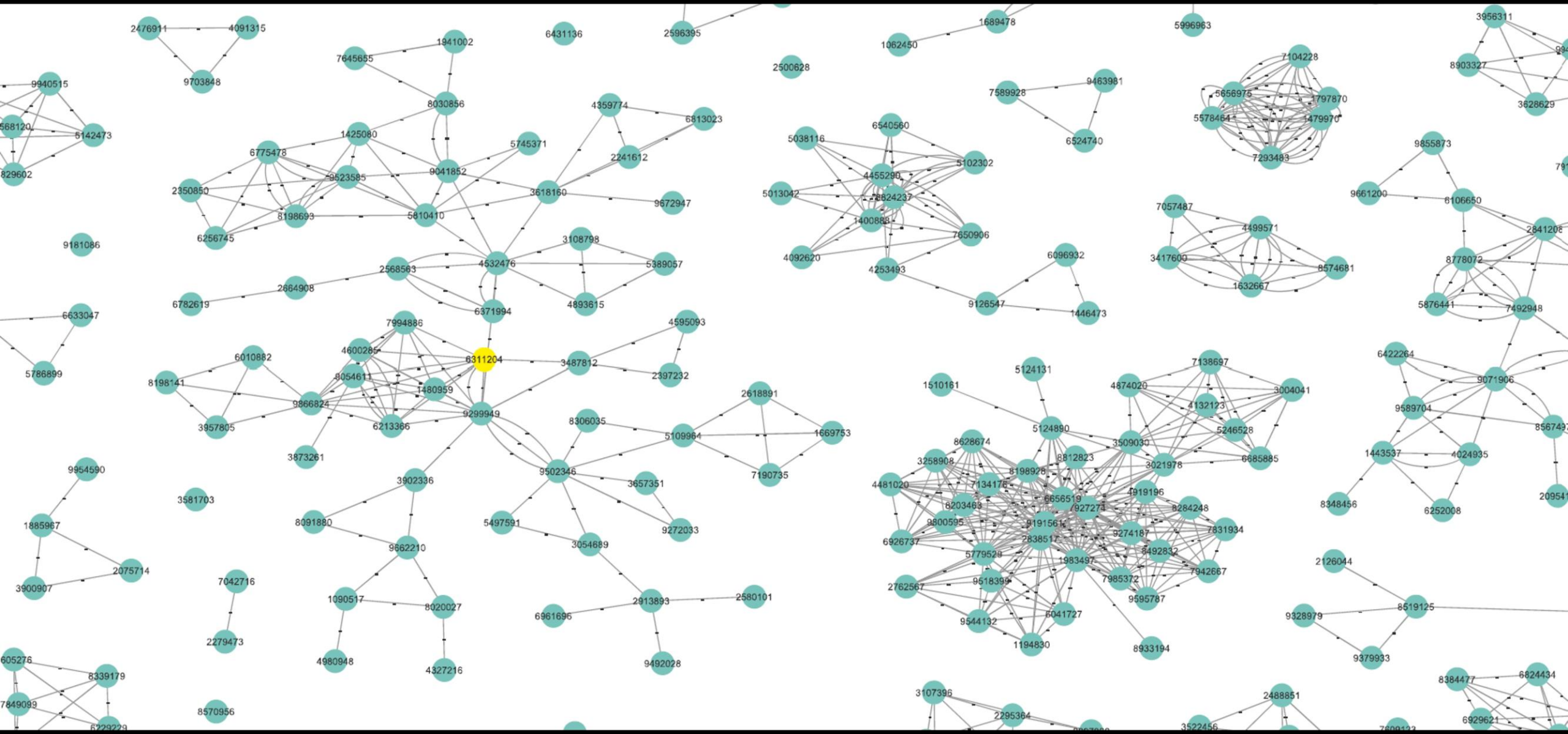


OSTI NFC: [’81-’82][’90][15,471][36,282][52,779]

- Training Data
  - OSTI NFC documents from 1981 through 1982
  - 15,471 authors
  - 36,282 edges in coauthorship network
  - 8,846 manuscript abstracts

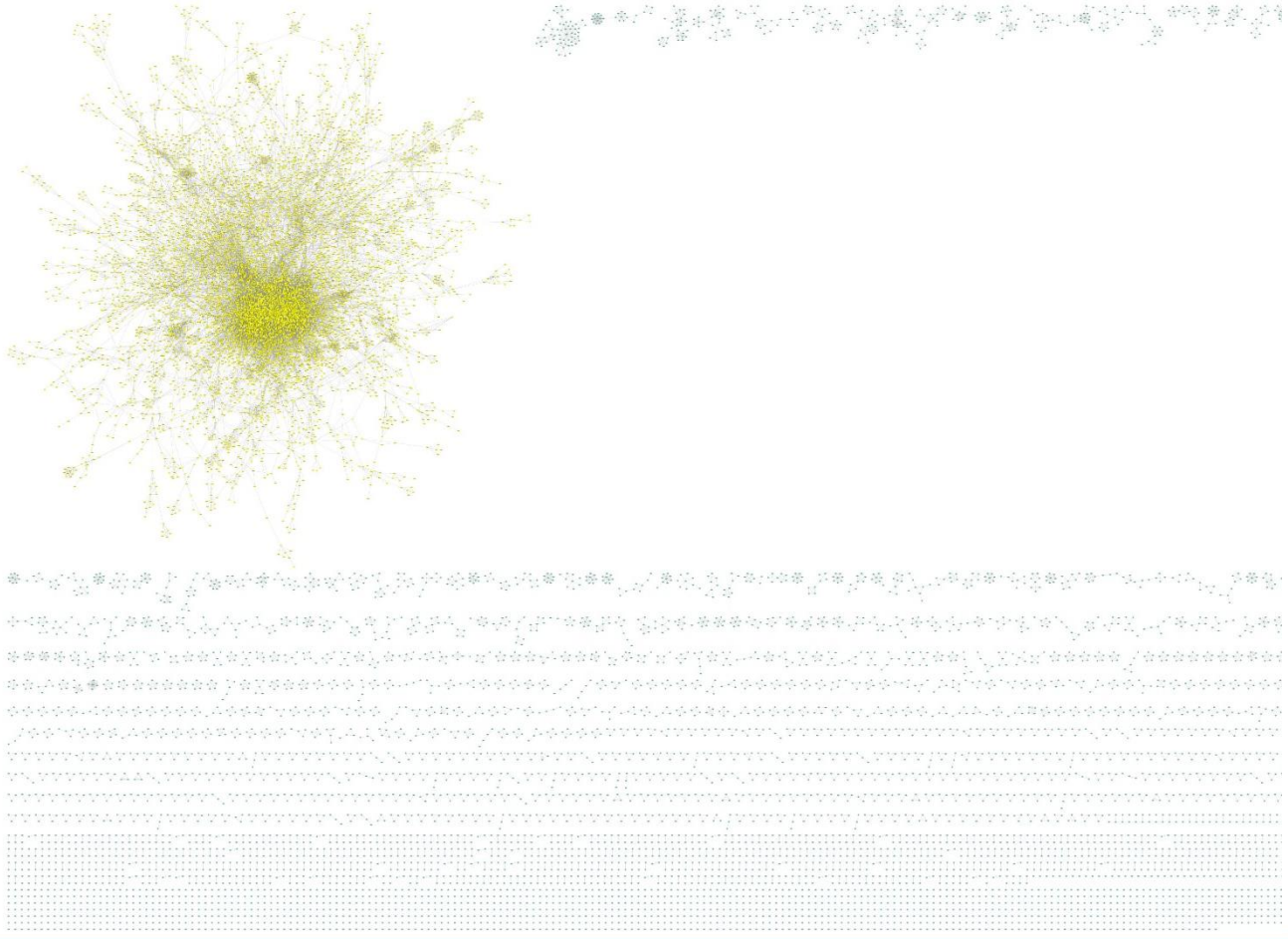
Property Name	Value
Avg. Number of Neighbors	8.602
Network Diameter	29
Clustering Coefficient	0.767
Network Density	0.002
Connected Components	4251




2476911
4091315
9703848
6431136
2596395
1941002
7645655
1689478
1062450
5996963
3956311
8903327
3628629
2500628
9940515
5142473
8030856
4359774
6813023
7589928
9463981
7104228
5656975
1797870
5578464
1479970
7293483
6524740
1425080
5745371
6775478
2241612
5038116
6540560
5102302
4455290
9523585
9041852
2350850
3618160
9672947
5013042
2824237
1400883
7650906
8198693
5810410
6256745
9181086
3108798
4092620
4253493
6096932
9126547
1446473
2568563
4532476
5389057
2664908
6782619
4893615
6371994
6633047
5786899
7994886
4595093
6010882
4600285
6311204
3487812
2397232
8198141
8054611
1480959
9866824
9299949
2618891
3957805
6213366
8306035
5109964
1669753
3873261
7190735
9954590
9502346
3657351
3902336
3581703
8091880
5497591
9272033
1885967
9662210
3054689
2075714
3900907
7042716
1090517
8020027
2913893
2580101
6961696
2279473
4980948
4327216
9492028
8339179
8570956
9855873
9661200
6106650
7057487
4499571
3417600
8574681
1632667
8778072
5876441
7492948
7138697
6422264
1510181
5124131
4874020
4132123
3004041
9071906
9589704
5124890
5246528
8628674
3509030
3258908
8812823
3021978
6685885
1443537
4024935
8198926
4481020
7134176
6656519
4919196
8203463
927273
8284248
8348456
6252008
9800595
9191561
9274187
7831934
6926737
2838517
8492832
5779528
1983497
7942667
2126044
2762567
9518399
7985372
9595787
8519125
9328979
9544132
6041727
1194830
8933194
9379933
8384477
6824434
3107396
2295364
2488851
6929621
3522456

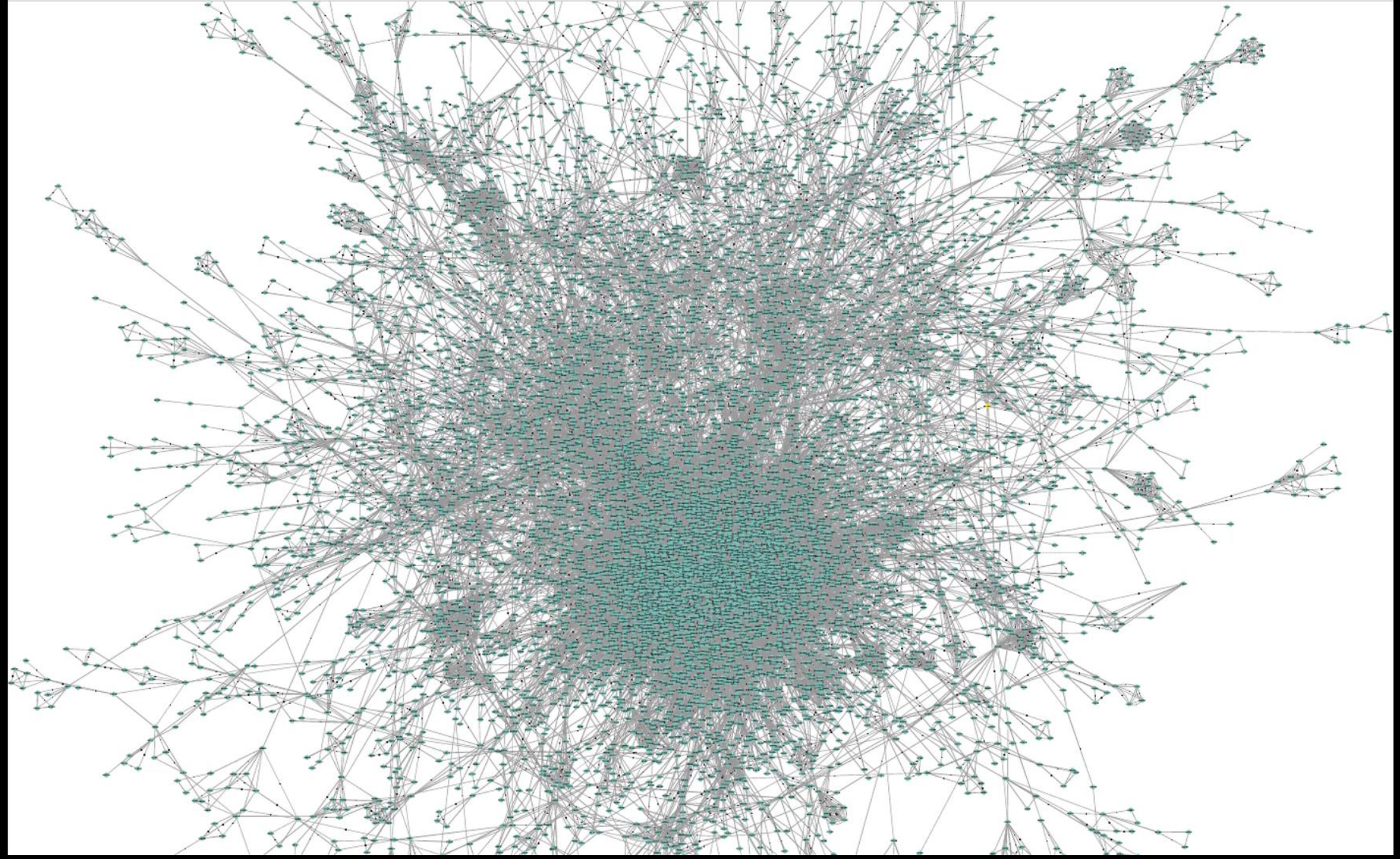
OSTI NFC: [’81-’82][’90][15,471][36,282][52,779]

- Test Data
  - OSTI NFC documents from 1983 through 1990
  - Restricted to same 15,471 authors
  - 52,779 – 36,282 = 16,497 new edges
  - 29,100 new manuscript abstracts



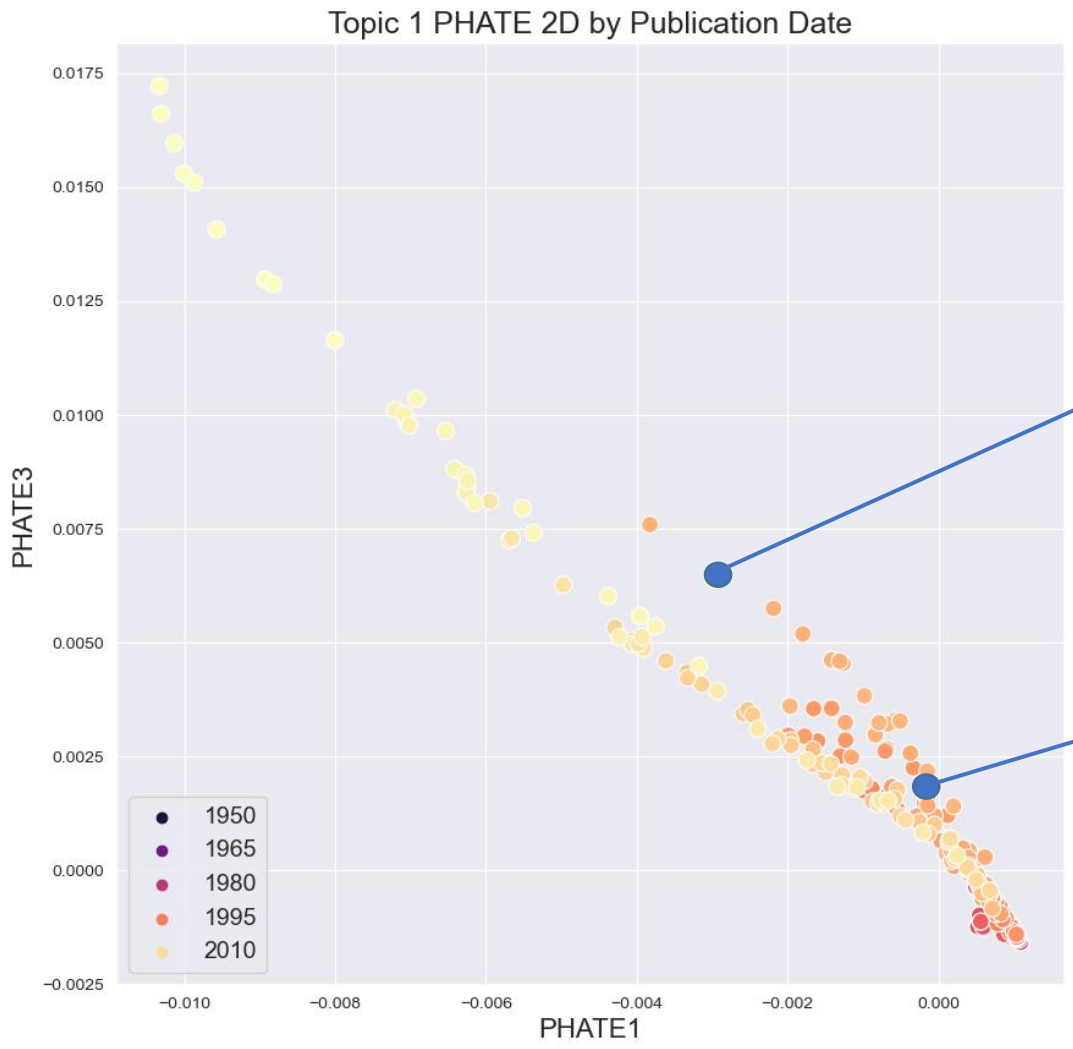
Property Name	Value
Avg. Number of Neighbors	8.864
Network Diameter	25
Clustering Coefficient	0.615
Network Density	0.001
Connected Components	3119

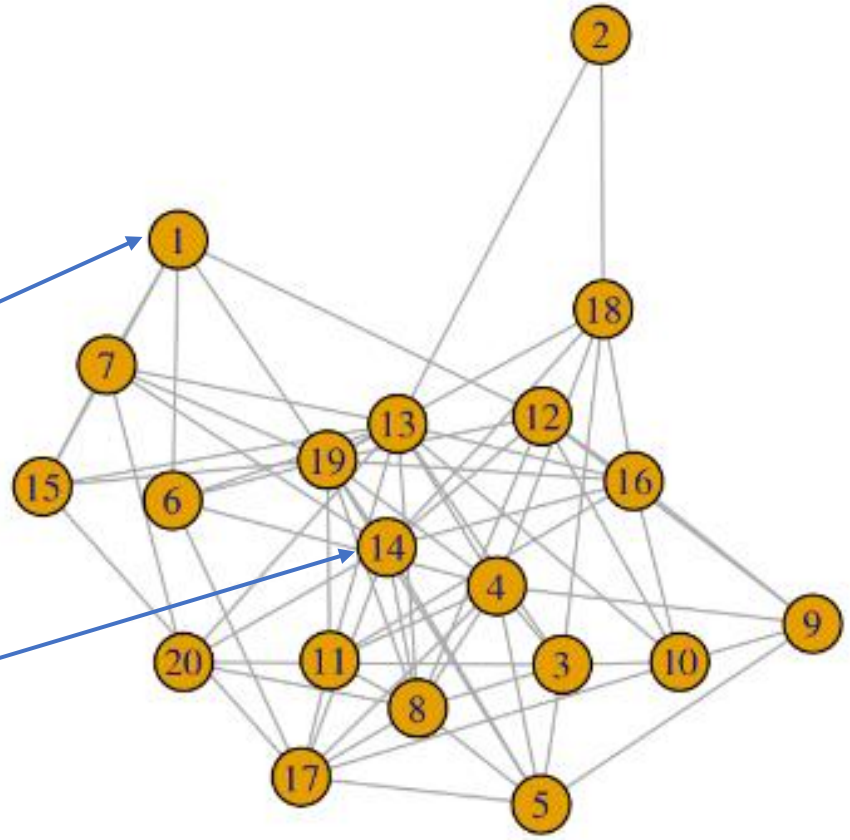
0



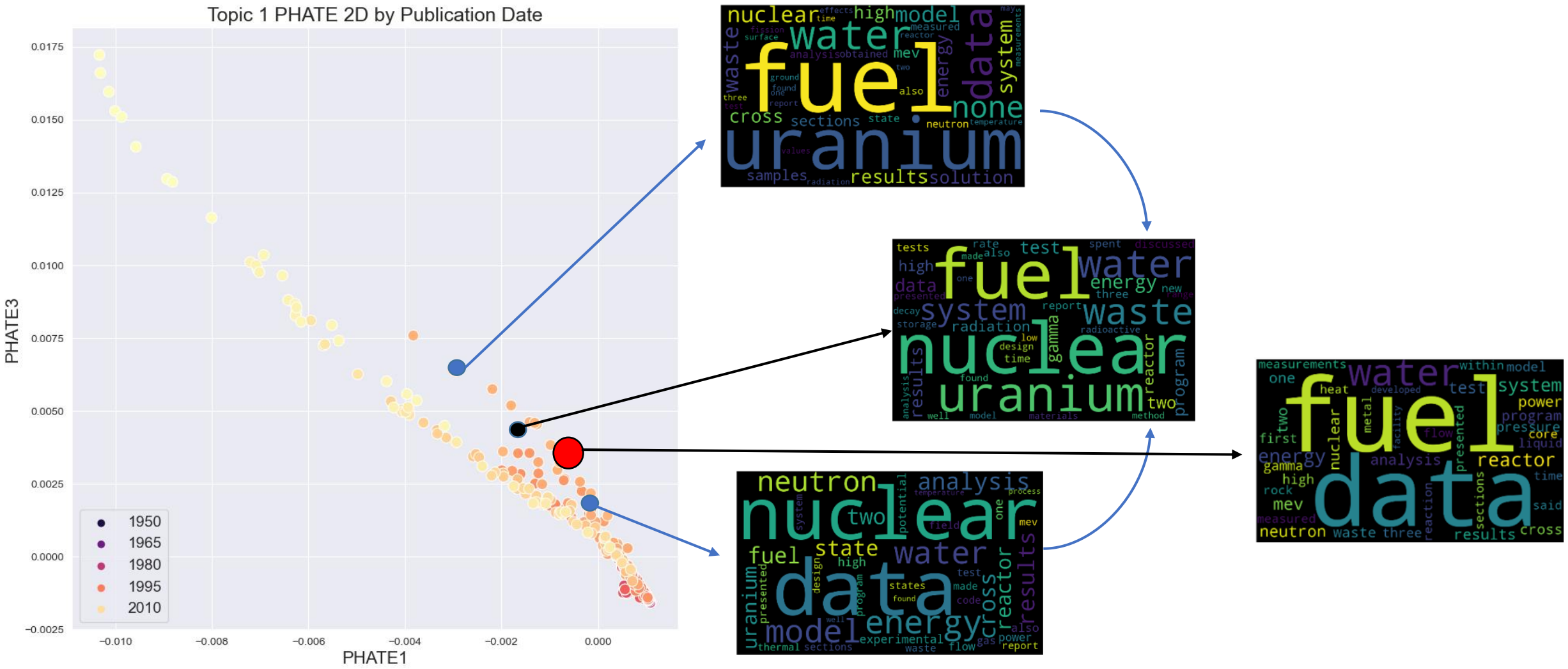
OSTI NFC: [’81-’82][’90][15,471][36,282][52,779]

Baseline Method	TP	FP	FN	Precision	Recall	F1-Score
Graph Distance 2 (GD2)	426	17,715	16,071	0.0235	0.0258	0.0246
Common Neighbors (CN)	**939**	**17,202**	**15,558**	**0.0518**	**0.0569**	**0.0542**
Adamic-Adar (AA)	905	17,236	15,592	0.0499	0.0549	0.0523
Jaccard Coefficient (JC)	619	17,522	15,878	0.0341	0.0375	0.0357
Resource Allocation (RA)	774	17,367	15,723	0.0427	0.0469	0.0447
AToMS Method	**TP**	**FP**	**FN**	**Precision**	**Recall**	**F1-Score**
Common Neighbors/Hellinger	**944**	**17,197**	**15,553**	**0.0520**	**0.0572**	**0.0545**
Adamic-Adar/Hellinger	907	17,234	15,590	0.0500	0.0550	0.0524
Jaccard Coefficient/Hellinger	527	17,614	15,970	0.0291	0.0319	0.0304
Resource Allocation/Hellinger	781	17,360	15,716	0.0431	0.0473	0.0450



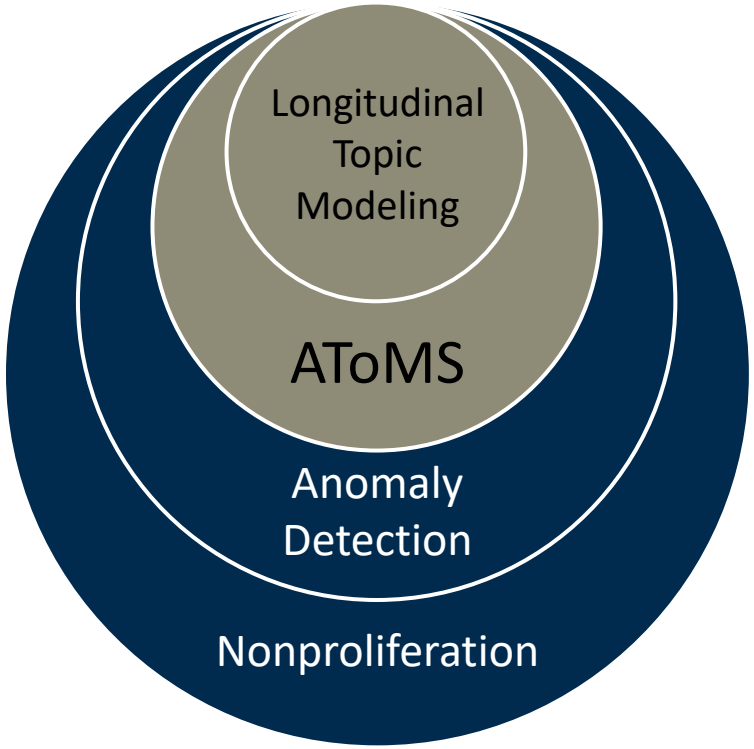


PNNL-SA-181770


Topic 1 PHATE 2D by Publication Date
PHATE3
PHATE1
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
−0.0025
−0.010
−0.008
−0.006
−0.004
−0.002
0.000
1950
1965
1980
1995
2010
fuel
uranium
water
nuclear
data
model
system
waste
energy
none
cross
sections
results
solution
samples
neutron
analysis
reactor
two

Conclusions and Future Work

- Link Prediction
  - Challenging problem in bibliographic networks
    - Networks are extremely sparse
  - AToMS outperforms “network-only” algorithms
    - Common neighbors is surprisingly effective
- Interpretability
  - AToMS provides predictive document modeling
  - Topic modeling reveals author interests
- Longitudinal Author-Topic Modeling
  - AToMS can be improved by using temporal information
  - Combine GDLTM longitudinal topics with AToMS model
- Nonproliferation Objective
  - Computational tools assist understanding bibliographic data
  - Can signal knowledge-based threats



ACKNOWLEDGEMENTS

This material is based upon work supported by the Department of Energy / National Nuclear Security Administration under Award Number(s) DE-NA0003921.

Georgia Tech

Clafflin University Founded 1869

Colorado School of Mines

Duke University

University of Hawai'i

MIT

University of Michigan

The University of North Carolina at Chapel Hill

The Ohio State University

Spelman College — A Choice to Change the World

Texas A&M University

The University of Texas at Austin

Washington State University

Argonne National Laboratory

Berkeley Lab

Brookhaven National Laboratory

INL Idaho National Laboratory — 70 Years of Science & Innovation

Lawrence Livermore National Laboratory

Los Alamos National Laboratory — EST. 1943

Nevada National Security Site — Managed and operated by Mission Support and Test Services

Oak Ridge National Laboratory

Pacific Northwest National Laboratory

PPPL Princeton Plasma Physics Laboratory

Sandia National Laboratories

SRNL Savannah River National Laboratory

National Nuclear Security Administration

Consortium for Enabling Technologies & Innovation

OSTI NFC 1.1 (897 Authors, 1275 edges, 560 edge budget)

Method Name	TP	FP	TN	FN	Precision	Recall	F1-Score
AToMSv2_pa_he	0	560	-	155	0.0000	0.0000	0.0000
AToMSv2_aa_he	16	544	-	139	0.0286	0.1032	0.0448
AToMSv2_jc_he	**20**	540	-	**135**	**0.0357**	**0.1290**	**0.0559**
AToMSv2_ra_he	17	543	-	138	0.0304	0.1097	0.0476
AToMSv2_cn_he	15	545	-	140	0.0268	0.0968	0.0420
AToMSv2_jc_co	15	545	-	140	0.0268	0.0968	0.0420
PA	1	559	-	154	0.0018	0.0065	0.0028
AA	15	545	-	140	0.0268	0.0968	0.0420
JC	16	544	-	139	0.0286	0.1032	0.0448
RA	16	544	-	139	0.0286	0.1032	0.0448
CN	15	545	-	140	0.0268	0.0968	0.0420
GD2	4	**152**	-	151	0.0256	0.0258	0.0257