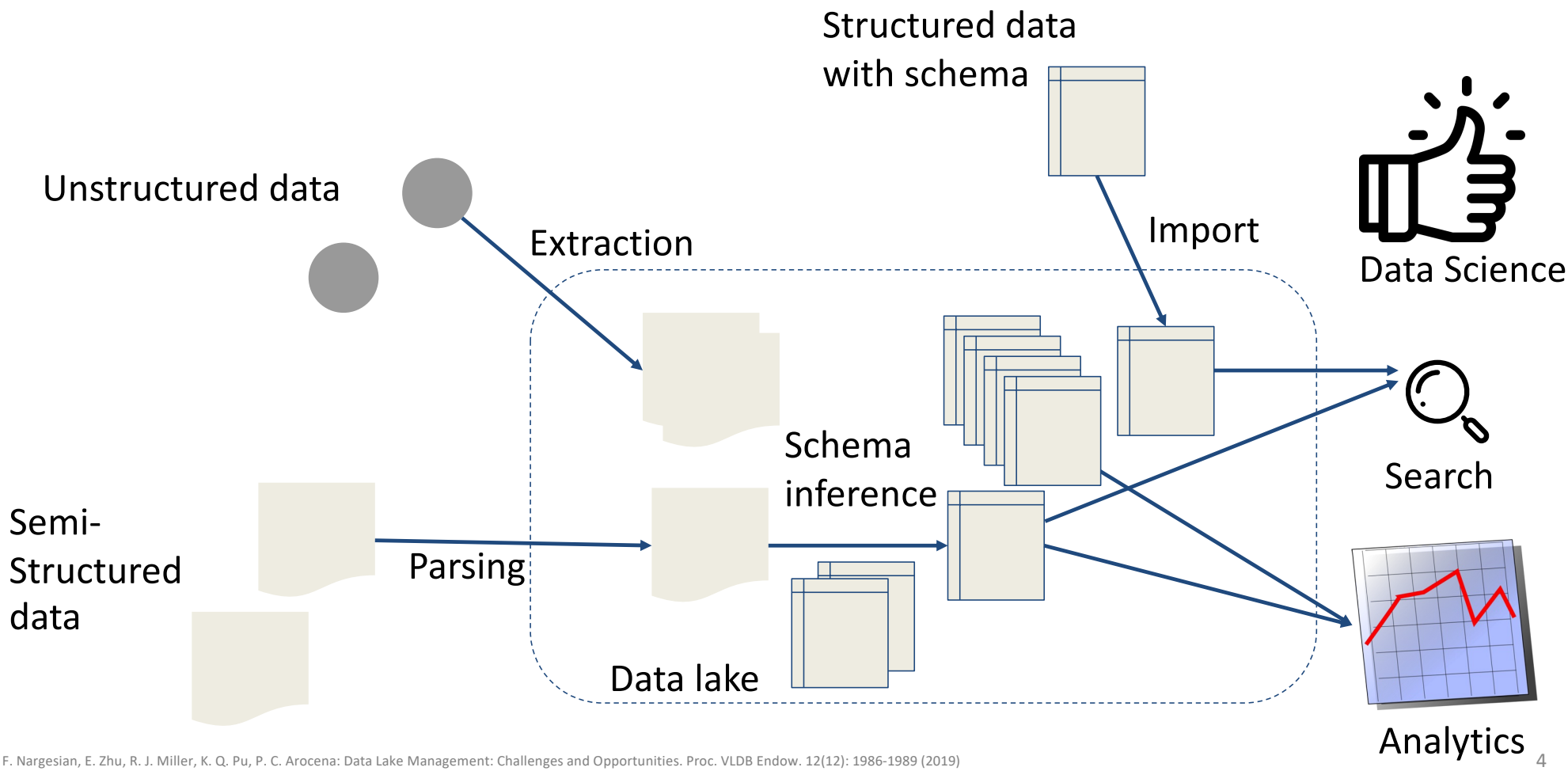
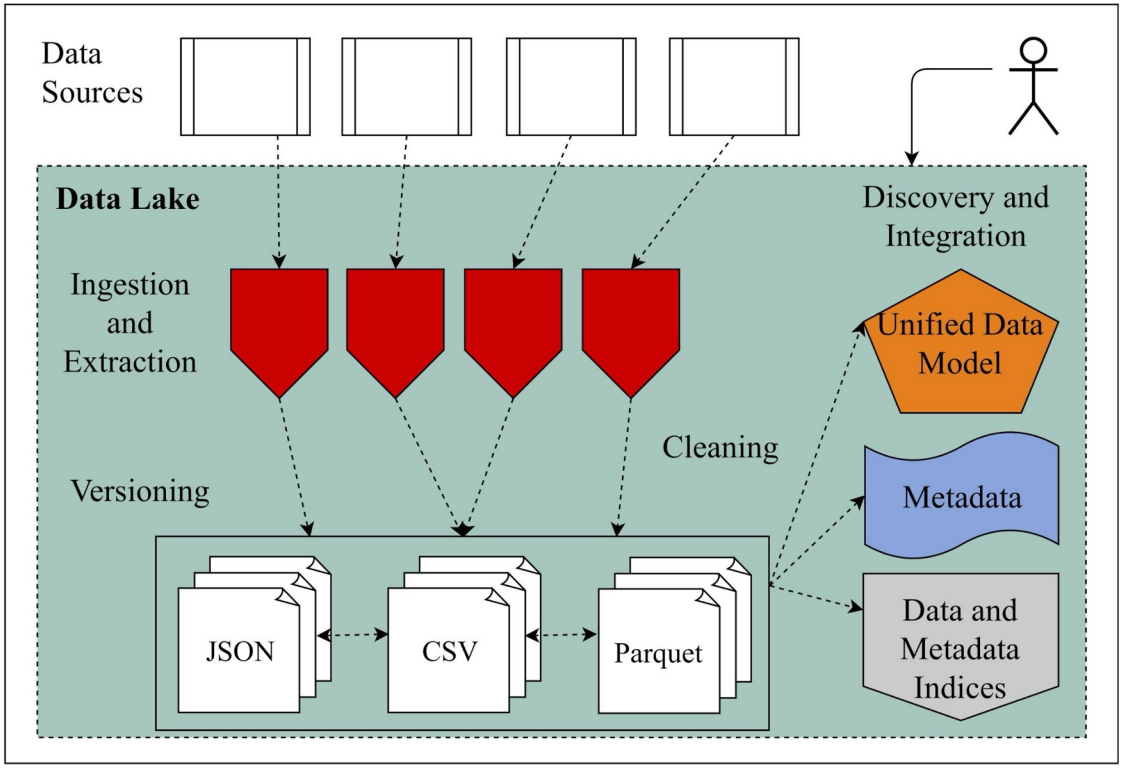
Data Lakes



F. Nargesian, E. Zhu, R. J. Miller, K. Q. Pu, P. C. Arocena: Data Lake Management: Challenges and Opportunities. Proc. VLDB Endow. 12(12): 1986-1989 (2019)

Common Tasks in Data Lakes

- Ingestion
- Extraction (Type Inference)
- Metadata Management
- Cleaning
- Integration
- Search/Discovery
- Versioning



F. Nargesian, E. Zhu, R. J. Miller, K. Q. Pu, P. C. Arocena: Data Lake Management: Challenges and Opportunities. Proc. VLDB Endow. 12(12): 1986-1989 (2019)

Data Lake Research Themes

- Search
  - Using metadata – Google dataset search [Brickley+19]
  - Using data
    - Relevant Table Search [Sarma+12]
    - Join Search [Zhu+16,Zhu+19]
    - Union Search [Nargesian+18]
- Metadata Discovery
  - Data Lake Organization [Nargesian+20]
  - Semantic Type Discovery [Hulsebos+19,Zhang+20,Ota+20]
  - Harmonization – encourage reuse of metadata and data definitions

In data lakes, before you can talk about integration, you need effective ways of finding the right data to solve a given data science problem.

D. Brickley, Matthew Burgess, Natasha F. Noy: Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. WWW 2019: 1365-1375
A. D. Sarma, L, Fang, N. Gupta, A. Y. Halevy, H. Lee, F. Wu, R. Xin, C. Yu: Finding related tables. SIGMOD Conference 2012: 817-828
E. Zhu, F. Nargesian, K. Q. Pu, R. J. Miller: LSH Ensemble: Internet-Scale Domain Search. Proc. VLDB Endow. 9(12): 1185-1196 (2016)
E. Zhu, D. Deng, F. Nargesian, R. J. Miller: JOSIE: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes. SIGMOD Conference 2019: 847-864
F. Nargesian, E. Zhu, K. Q. Pu, R. J. Miller: Table Union Search on Open Data. Proc. VLDB Endow. 11(7): 813-825 (2018)
F. Nargesian, K. Q. Pu, E. Zhu, B. G. Bashardoost, R. J. Miller. Orgganizing Data Lakes for Navigation. SIGMOD Conference 2020: 1939-1950
M. Hulsebos, K. Z. Hu, M. A. Bakker, E. Zgraggen, A. Satyanarayan, T. Kraska, Ç. Demiralp, and C. A. Hidalgo. Sherlock: A deep learning approach to semantic data type detection. In SIGKDD, pages 1500–1508. ACM, 2019
D. Zhang, Y. Suhara, J. Li, M. Hulsebos, Ç. Demiralp, and W. Tan. Sato: Contextual semantic type detection in tables. PVLDB, 13(11):1835–1848, 2020
M. Ota, H. Mueller, J. Freire, and D. Srivastava. Data-driven domain discovery for structured datasets. PVLDB, 13(7):953–965, 2020.

Theme for talk

- Data lakes present both new challenges and new opportunities for solving complex data integration problems
- To make this point, I'm going to present some new work from EDBT 2021
  - View data lake as a network
  - Use network centrality measures as a way of extracting hidden semantics behind values in a data lake
- Network science view of data lakes
  - We can **apply** network science abstractions to better understand data lakes
  - But can we also use data lakes to **advance** network science?

Data Lakes

Data Lakes are deeply heterogenous where the same data value can have multiple meanings

An Example Data Lake

Data Lakes are deeply heterogenous where the same data value can have multiple meanings

Col1	Col2	Col3
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

name_1	name_2	name_3
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

An Example Data Lake

Data Lakes are deeply heterogenous where the same data value can have multiple meanings

Col1	Col2	Col3
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

US State Abbreviations?

name_1	name_2	name_3
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

An Example Data Lake

Data Lakes are deeply heterogenous where the same data value can have multiple meanings

Col1	Col2	Col3
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

Years?

name_1	name_2	name_3
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

An Example Data Lake

Data Lakes are deeply heterogenous where the same data value can have multiple meanings

Col1	Col2	Col3
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

City in California?

name_1	name_2	name_3
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

An Example Data Lake

Data Lakes are deeply heterogenous where the same data value can have multiple meanings

Books

Title	Year	Pages
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

name_1	name_2	name_3
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

An Example Data Lake

Data Lakes are deeply heterogenous where the same data value can have multiple meanings

Books

Title	Year	Pages
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

Locations

Location	US State	Population
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

An Example Data Lake

Data Lakes are deeply heterogenous where the same data value can have multiple meanings

Books

Title	Year	Pages
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

Locations

Location	US State	Population
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

Airlines

	Airline	Flight	Departure
Air China	CA	129	12:50
Malév Hungarian Airlines	MA	355	16:45
Air Madrid	NM	280	7:35

An Example Data Lake

Data Lakes are deeply heterogenous where the same data value can have multiple meanings

Books

Title	Year	Pages
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

Locations

Location	US State	Population
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

California

Airlines

Airline	Flight	Departure
CA	129	12:50
MA	355	16:45
NM	280	7:35

Air China

Malév Hungarian Airlines

Air Madrid

An Example Data Lake

Data Lakes are deeply heterogenous where the same data value can have multiple meanings

Books

Title	Year	Pages
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

Novel

County

Locations

Location	US State	Population
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

Airlines

Airline	Flight	Departure
Air China	CA	129	12:50
Malév Hungarian Airlines	MA	355	16:45
Air Madrid	NM	280	7:35

An Example Data Lake

Homograph: A data value in the data lake with more than one meaning

Books

Title	Year	Pages
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

Locations

Location	US State	Population
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

Airlines

Airline	Flight	Departure
CA	129	12:50
MA	355	16:45
NM	280	7:35

Disambiguation in Data Lakes

F. Nargesian, E. Zhu, K. Q. Pu, and R. J. Miller. Table union search on open data. PVLDB, 11(7):813–825, 2018.
M. Hulsebos, K. Z. Hu, M. A. Bakker, E. Zgraggen, A. Satyanarayan, T. Kraska, Ç. Demiralp, and C. A. Hidalgo. Sherlock: A deep learning approach to semantic data type detection. In SIGKDD, pages 1500–1508. ACM, 2019
D. Zhang, Y. Suhara, J. Li, M. Hulsebos, Ç. Demiralp, and W. Tan. Sato: Contextual semantic type detection in tables. PVLDB, 13(11):1835–1848, 2020
M. Ota, H. Mueller, J. Freire, and D. Srivastava. Data-driven domain discovery for structured datasets. PVLDB, 13(7):953–965, 2020.

Disambiguation in Data Lakes

Supervised

- Knowledge-based Techniques
  - Map data values to DBpedia or YAGO
  - Low coverage in data lakes [Nargesian+ 2018]
- ML Techniques
  - Sherlock [Hulsebos+ 2019], SATO [Zhang+ 2020]
  - Trained on only 78 types!

F. Nargesian, E. Zhu, K. Q. Pu, and R. J. Miller. Table union search on open data. PVLDB, 11(7):813–825, 2018.
M. Hulsebos, K. Z. Hu, M. A. Bakker, E. Zgraggen, A. Satyanarayan, T. Kraska, Ç. Demiralp, and C. A. Hidalgo. Sherlock: A deep learning approach to semantic data type detection. In SIGKDD, pages 1500–1508. ACM, 2019
D. Zhang, Y. Suhara, J. Li, M. Hulsebos, Ç. Demiralp, and W. Tan. Sato: Contextual semantic type detection in tables. PVLDB, 13(11):1835–1848, 2020
M. Ota, H. Mueller, J. Freire, and D. Srivastava. Data-driven domain discovery for structured datasets. PVLDB, 13(7):953–965, 2020.

Disambiguation in Data Lakes

Supervised

- Knowledge-based Techniques
  - Map data values to DBpedia or YAGO
  - Low coverage in data lakes [Nargesian+ 2018]
- ML Techniques
  - Sherlock [Hulsebos+ 2019], SATO [Zhang+ 2020]
  - Trained on only 78 types!

Unsupervised

- Unsupervised Domain Discovery
  - D^4 [Ota+ 2020]
  - Groups data values to domains
  - If a data value placed in more than one domain, it may be a homograph

F. Nargesian, E. Zhu, K. Q. Pu, and R. J. Miller. Table union search on open data. PVLDB, 11(7):813–825, 2018.
M. Hulsebos, K. Z. Hu, M. A. Bakker, E. Zgraggen, A. Satyanarayan, T. Kraska, Ç. Demiralp, and C. A. Hidalgo. Sherlock: A deep learning approach to semantic data type detection. In SIGKDD, pages 1500–1508. ACM, 2019
D. Zhang, Y. Suhara, J. Li, M. Hulsebos, Ç. Demiralp, and W. Tan. Sato: Contextual semantic type detection in tables. PVLDB, 13(11):1835–1848, 2020
M. Ota, H. Mueller, J. Freire, and D. Srivastava. Data-driven domain discovery for structured datasets. PVLDB, 13(7):953–965, 2020.

Disambiguation in Data Lakes

Supervised

- Knowledge-based Techniques
  - Map data values to DBpedia or YAGO
  - Low coverage in data lakes [Nargesian+ 2018]
- ML Techniques
  - Sherlock [Hulsebos+ 2019], SATO [Zhang+ 2020]
  - Trained on only 78 types!

Unsupervised

- Unsupervised Domain Discovery
  - D^4 [Ota+ 2020]
  - Groups data values to domains
  - If a data value placed in more than one domain, it may be a homograph

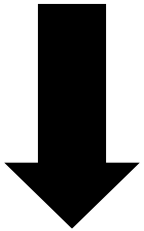
F. Nargesian, E. Zhu, K. Q. Pu, and R. J. Miller. Table union search on open data. PVLDB, 11(7):813–825, 2018.
M. Hulsebos, K. Z. Hu, M. A. Bakker, E. Zgraggen, A. Satyanarayan, T. Kraska, Ç. Demiralp, and C. A. Hidalgo. Sherlock: A deep learning approach to semantic data type detection. In SIGKDD, pages 1500–1508. ACM, 2019
D. Zhang, Y. Suhara, J. Li, M. Hulsebos, Ç. Demiralp, and W. Tan. Sato: Contextual semantic type detection in tables. PVLDB, 13(11):1835–1848, 2020
M. Ota, H. Mueller, J. Freire, and D. Srivastava. Data-driven domain discovery for structured datasets. PVLDB, 13(7):953–965, 2020.

Data Lake Disambiguation

Data Lake Disambiguation: Given a data lake containing a collection of tables determine if a data value that appears in more than one attribute or table has a single meaning or more than one meaning

Data Lake Disambiguation: DomainNet

Data Lake Disambiguation: Given a data lake containing a collection of tables determine if a data value that appears in more than one attribute or table has a single meaning or more than one meaning

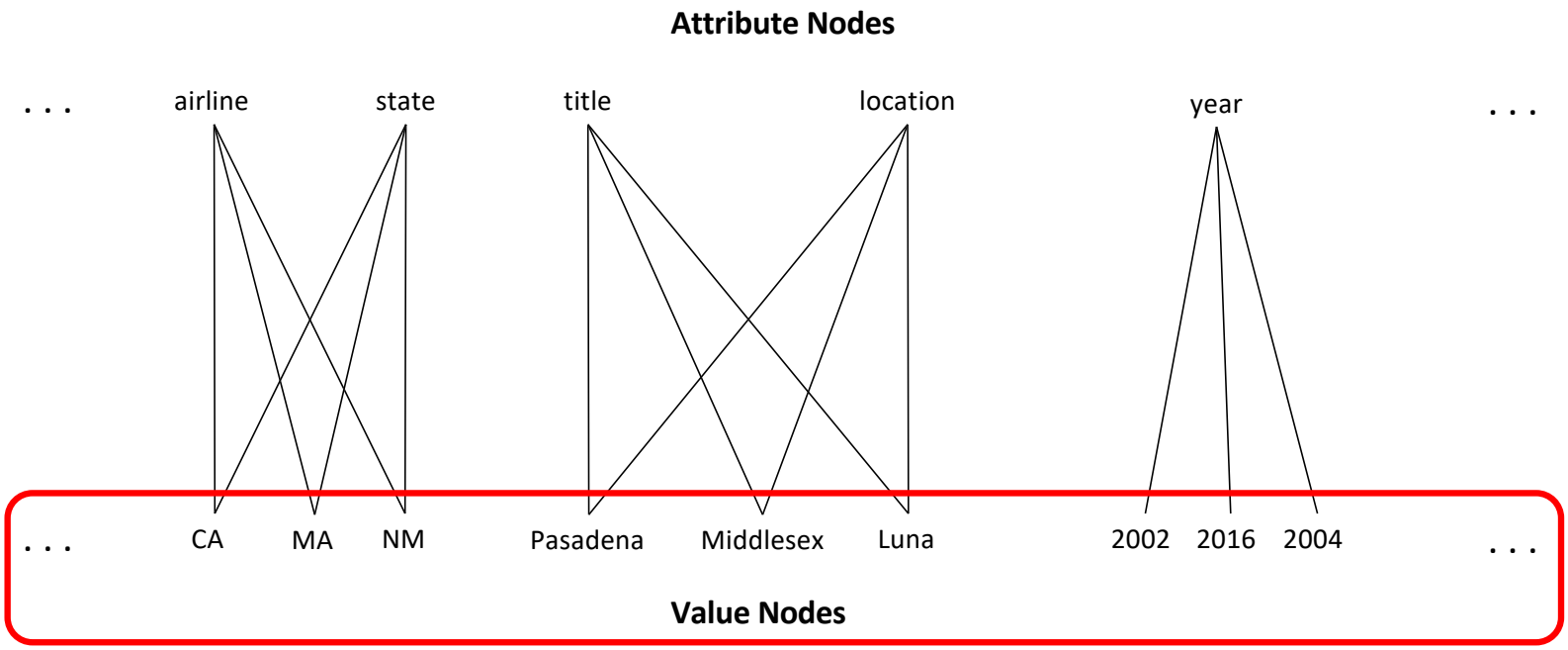


DomainNet: unsupervised technique to identify if a data value is a homograph or not

- Examines data value co-occurrence
- Uses a network-centrality measure on a bipartite graph representation of the data lake
- Given a homograph identify its number of meanings and group its attributes by their meaning

DomainNet (A Graph Representation)

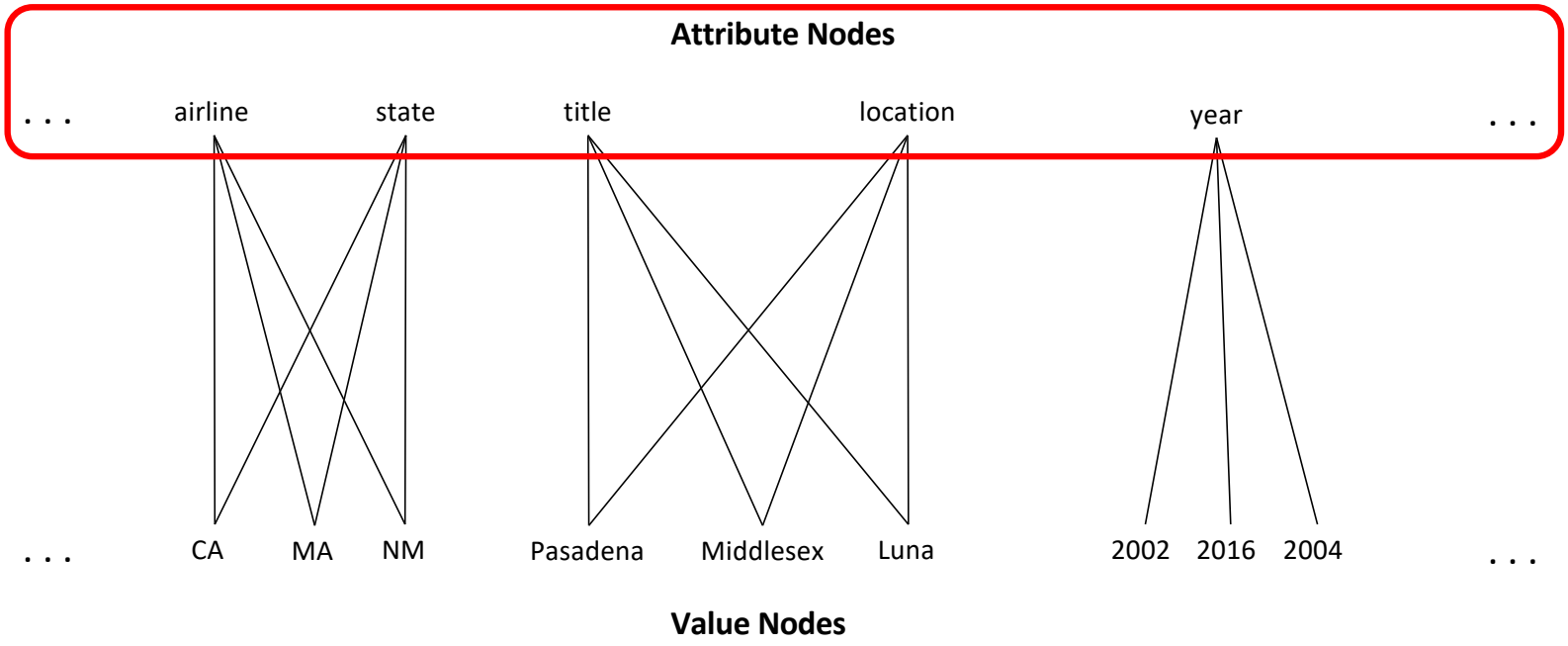
Intuition: A *homograph* likely co-occurs with a set of values that do not co-occur frequently with each other



Set semantics

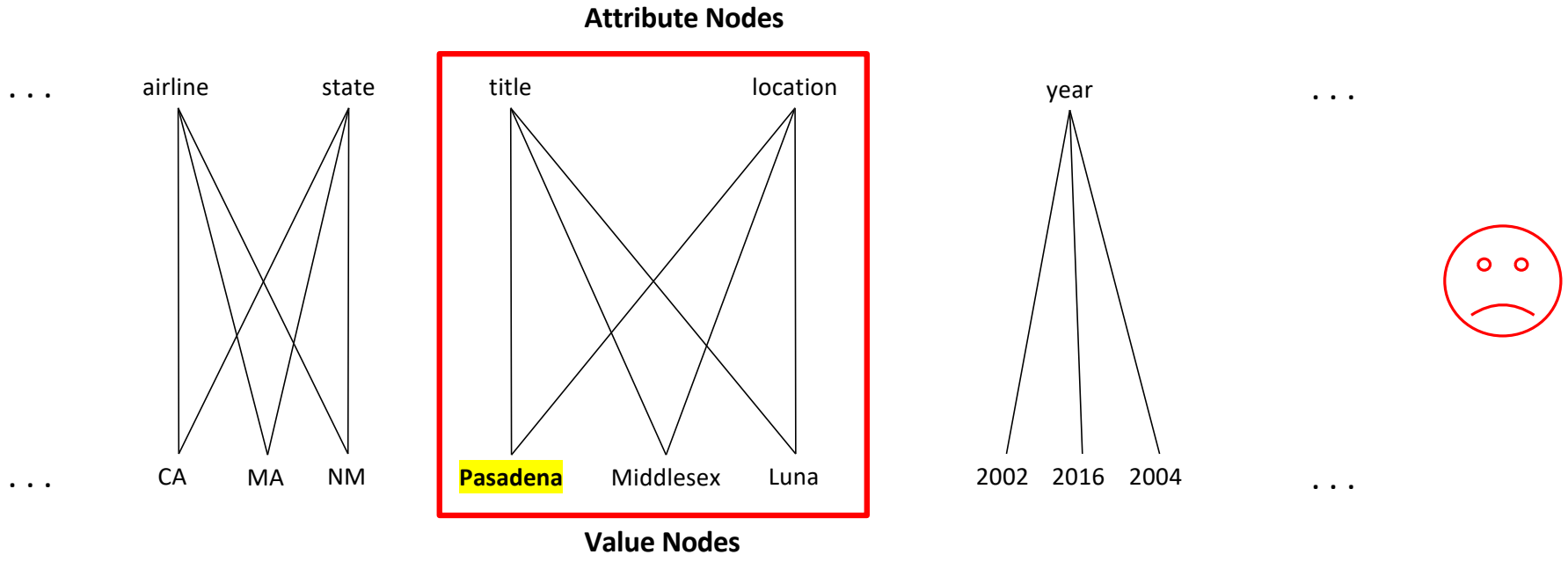
DomainNet (A Graph Representation)

Intuition: A *homograph* likely co-occurs with a set of values that do not co-occur frequently with each other



DomainNet (A Graph Representation)

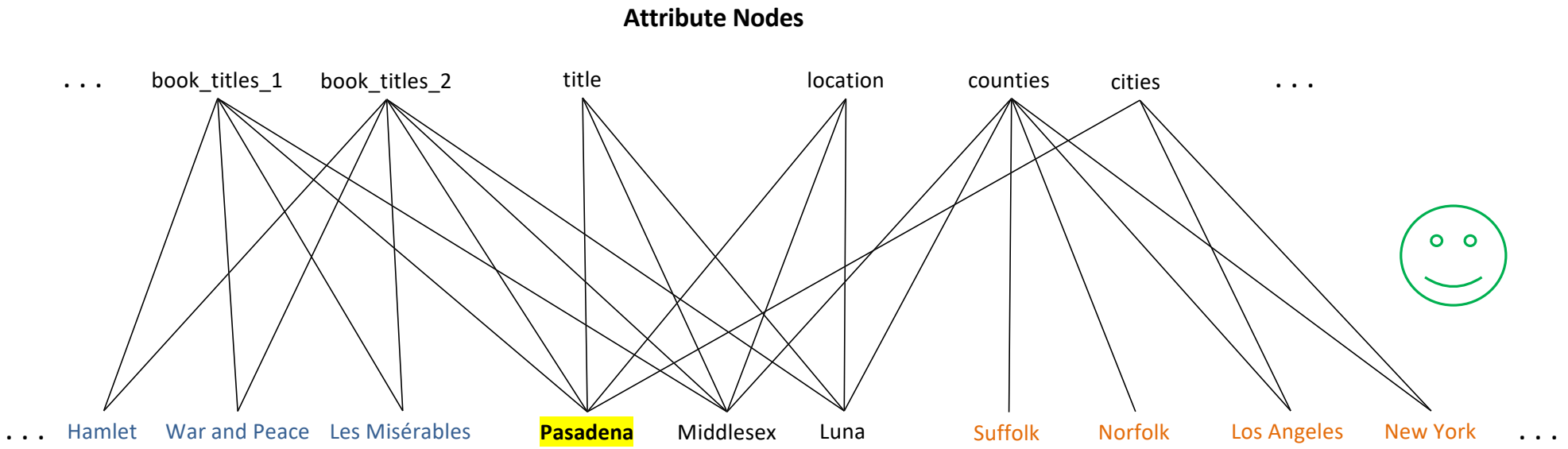
Intuition: A *homograph* likely co-occurs with a set of values that do not co-occur frequently with each other



DomainNet (A Graph Representation)

Intuition: A *homograph* likely co-occurs with a set of values that do not co-occur frequently with each other

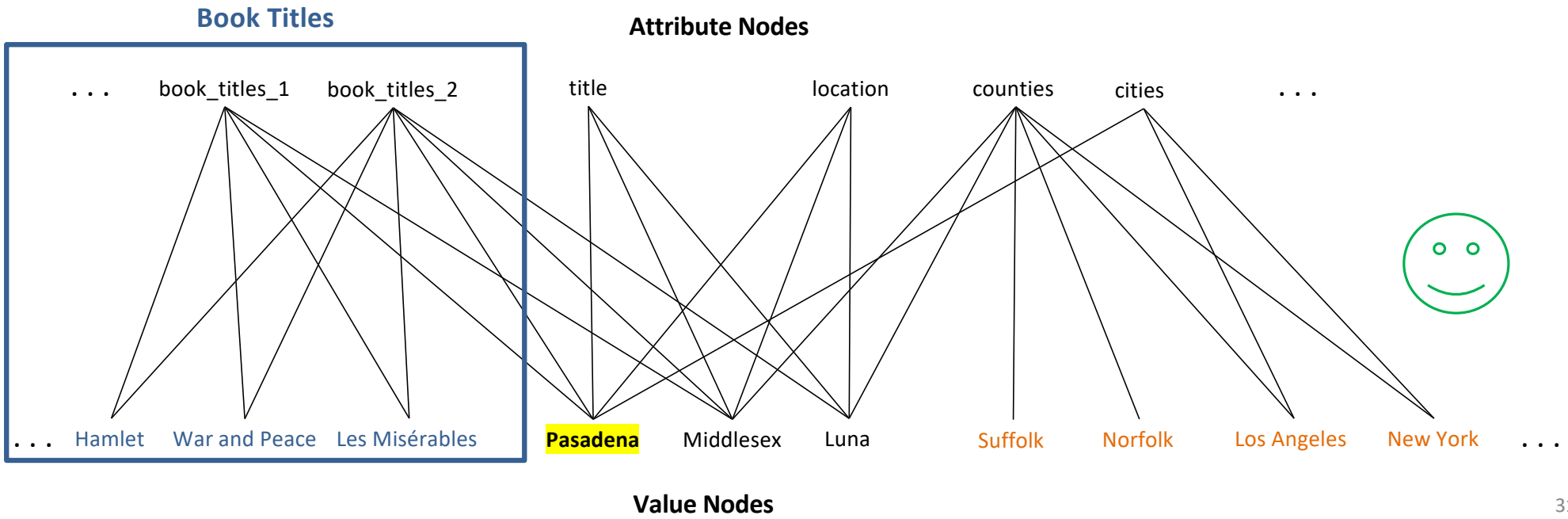
With enough data the co-occurrence pattern should emerge



DomainNet (A Graph Representation)

Intuition: A *homograph* likely co-occurs with a set of values that do not co-occur frequently with each other

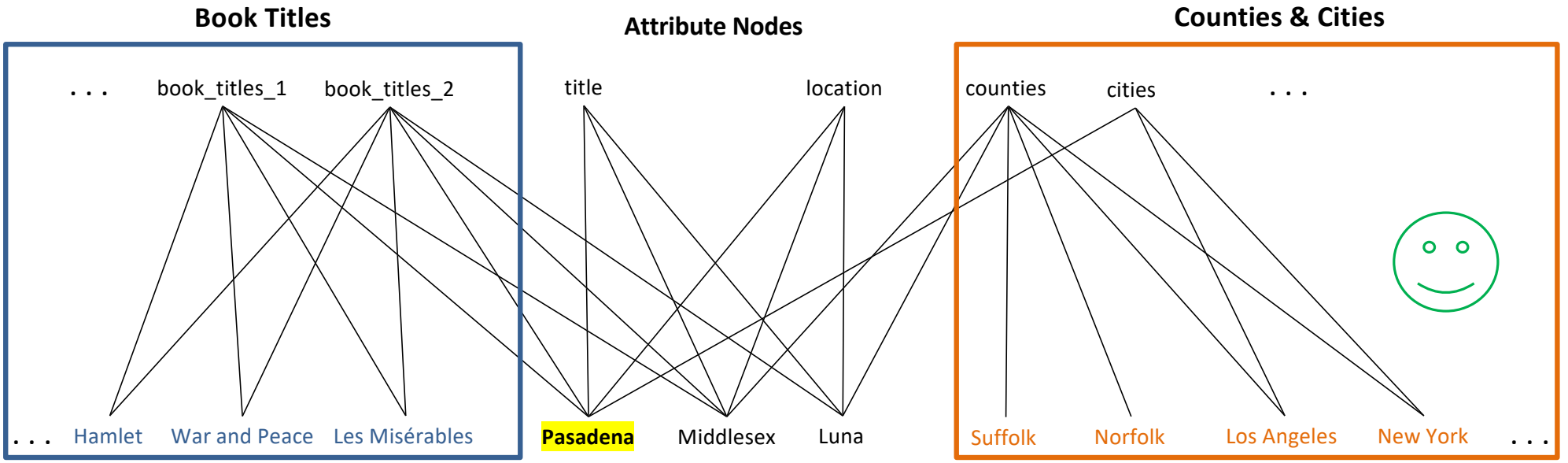
With enough data the co-occurrence pattern should emerge



DomainNet (A Graph Representation)

Intuition: A *homograph* likely co-occurs with a set of values that do not co-occur frequently with each other

With enough data the co-occurrence pattern should emerge



DomainNet (A Homograph Score)

Can we use the co-occurrence intuition to assign homograph scores for each data value?

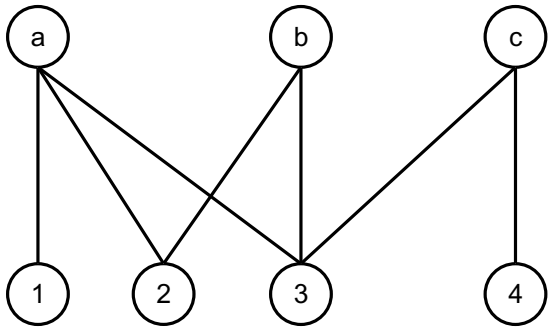
Considering only direct co-occurrences can be lossy!

Extract co-occurrence information over longer paths as well

We need a **global connectivity measure**!

DomainNet (Betweenness Centrality (BC))

- Betweenness centrality (BC) [Freeman 1977] of a node measures how often a node lies on shortest paths between all other nodes in the graph

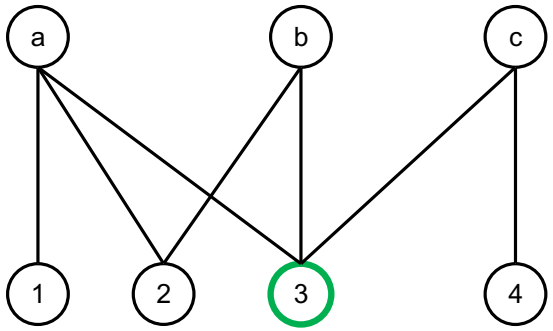


L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1977

Betweenness Centrality (BC)

- Betweenness centrality (BC) [Freeman 1977] of a node measures how often a node lies on shortest paths between all other nodes in the graph

$$\mathrm{BC}(3) = \sum_{v \neq 3,\, w \neq 3} \frac{\sigma_{vw}(3)}{\sigma_{vw}}$$



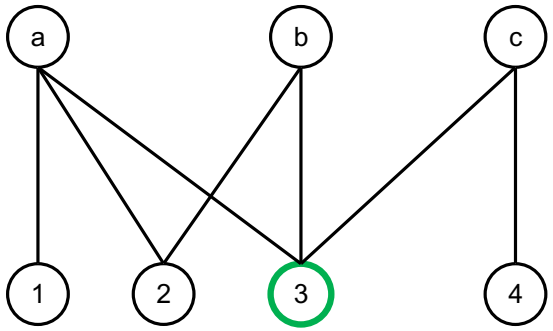
$\sigma_{vw}(u)$: number of shortest paths from v to w that pass through u
σ_{vw}: number of shortest paths from v to w

L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1977

Betweenness Centrality (BC)

- Betweenness centrality (BC) [Freeman 1977] of a node measures how often a node lies on shortest paths between all other nodes in the graph

$$\mathrm{BC}(3) = \sum_{v \neq 3,\, w \neq 3} \frac{\sigma_{vw}(3)}{\sigma_{vw}}$$



$\sigma_{vw}(u)$: number of shortest paths from v to w that pass through u

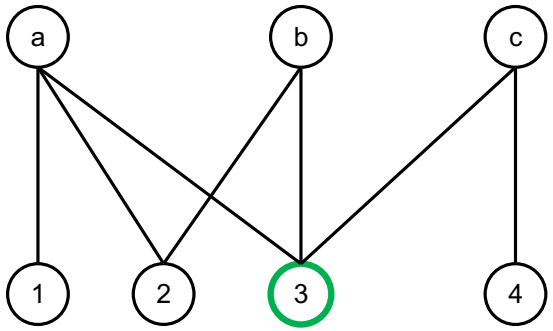
σ_{vw}: number of shortest paths from v to w

L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1977

Betweenness Centrality (BC)

- Betweenness centrality (BC) [Freeman 1977] of a node measures how often a node lies on shortest paths between all other nodes in the graph

$$\mathrm{BC}(3) = \sum_{v \neq 3,\, w \neq 3} \frac{\sigma_{vw}(3)}{\sigma_{vw}}$$



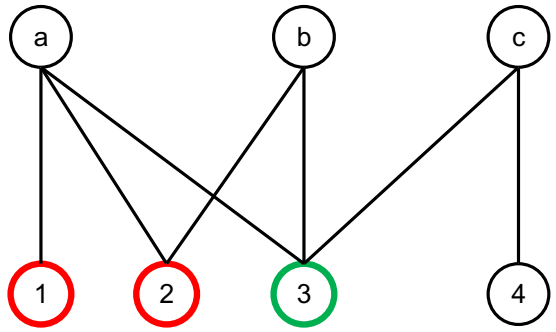
$\sigma_{vw}(u)$: number of shortest paths from v to w that pass through u
σ_{vw}: number of shortest paths from v to w

L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1977

Betweenness Centrality (BC)

- Betweenness centrality (BC) [Freeman 1977] of a node measures how often a node lies on shortest paths between all other nodes in the graph

$$\mathrm{BC}(3) = \sum_{v \neq 3,\, w \neq 3} \frac{\sigma_{vw}(3)}{\sigma_{vw}}$$



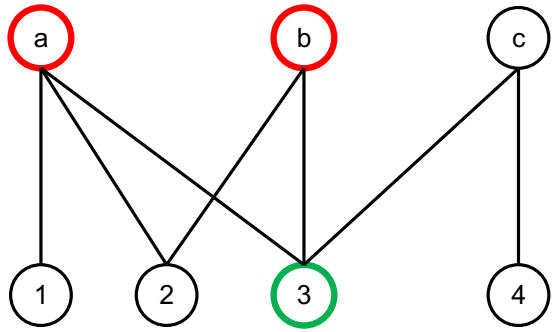
$\sigma_{vw}(u)$: number of shortest paths from v to w that pass through u
σ_{vw}: number of shortest paths from v to w

L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1977

Betweenness Centrality (BC)

- Betweenness centrality (BC) [Freeman 1977] of a node measures how often a node lies on shortest paths between all other nodes in the graph

$$\mathrm{BC}(3) = \sum_{v \neq 3,\, w \neq 3} \frac{\sigma_{vw}(3)}{\sigma_{vw}}$$



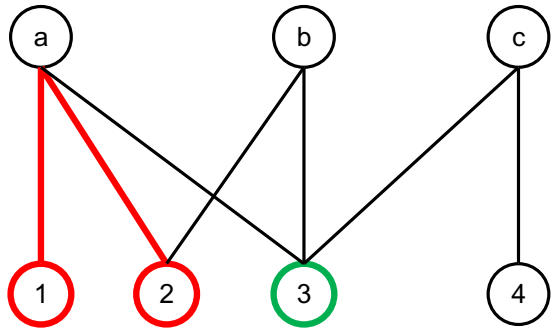
$\sigma_{vw}(u)$: number of shortest paths from v to w that pass through u
σ_{vw}: number of shortest paths from v to w

L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1977

Betweenness Centrality (BC)

- Betweenness centrality (BC) [Freeman 1977] of a node measures how often a node lies on shortest paths between all other nodes in the graph

$$\mathrm{BC}(3) = \sum_{v \neq 3,\, w \neq 3} \frac{\sigma_{vw}(3)}{\sigma_{vw}} = 0 + \cdots$$



$$\frac{\sigma_{1,2}(3)}{\sigma_{1,2}} = 0$$

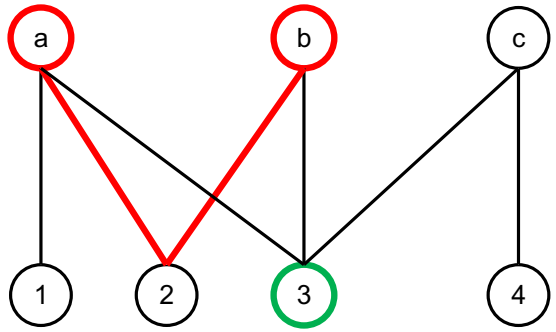
$\sigma_{vw}(u)$: number of shortest paths from v to w that pass through u
σ_{vw}: number of shortest paths from v to w

L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1977

Betweenness Centrality (BC)

- Betweenness centrality (BC) [Freeman 1977] of a node measures how often a node lies on shortest paths between all other nodes in the graph

$$\mathrm{BC}(3) = \sum_{v \neq 3,\, w \neq 3} \frac{\sigma_{vw}(3)}{\sigma_{vw}} = 0 + \cdots$$



$$\frac{\sigma_{1,2}(3)}{\sigma_{1,2}} = 0$$

Shortest paths between nodes *a* and *b*:

- a,2,b

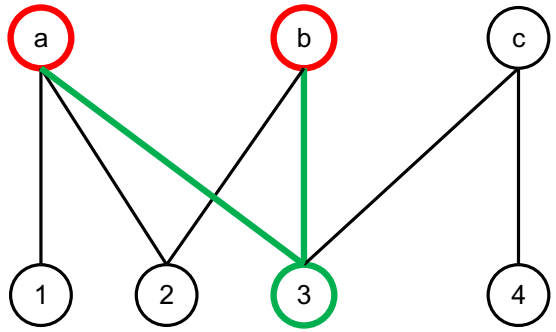
$\sigma_{vw}(u)$: number of shortest paths from v to w that pass through u
σ_{vw}: number of shortest paths from v to w

L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1977

Betweenness Centrality (BC)

- Betweenness centrality (BC) [Freeman 1977] of a node measures how often a node lies on shortest paths between all other nodes in the graph

$$\mathrm{BC}(3) = \sum_{v \neq 3,\, w \neq 3} \frac{\sigma_{vw}(3)}{\sigma_{vw}} = 0 + \cdots$$



$$\frac{\sigma_{1,2}(3)}{\sigma_{1,2}} = 0$$

Shortest paths between nodes *a* and *b*:

- a,2,b
- a,3,b

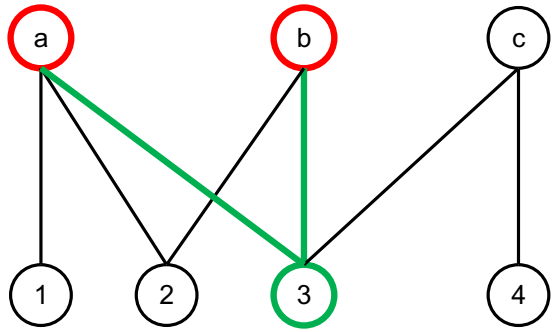
σ_{vw} are the total number of shortest paths from v to w and $\sigma_{vw}(u)$ are the number of shortest paths from v to w that pass through u

L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1977

Betweenness Centrality (BC)

- Betweenness centrality (BC) [Freeman 1977] of a node measures how often a node lies on shortest paths between all other nodes in the graph

$$\mathrm{BC}(3) = \sum_{v \neq 3,\, w \neq 3} \frac{\sigma_{vw}(3)}{\sigma_{vw}} = 0 + \frac{1}{2} + \cdots$$



$$\frac{\sigma_{1,2}(3)}{\sigma_{1,2}} = 0$$

Shortest paths between nodes *a* and *b*:

- a,2,b
- a,3,b

$$\frac{\sigma_{a,b}(3)}{\sigma_{a,b}} = \frac{1}{2}$$

σ_{vw} are the total number of shortest paths from v to w and $\sigma_{vw}(u)$ are the number of shortest paths from v to w that pass through u

L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1977

Betweenness Centrality (BC)

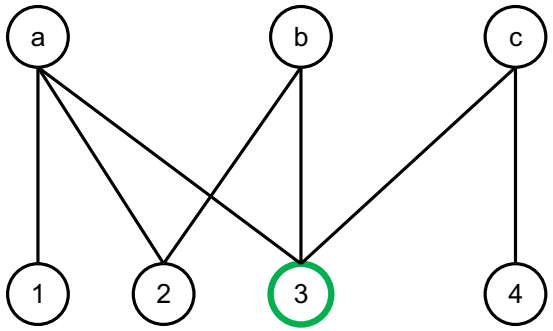
- Betweenness centrality (BC) [Freeman 1977] of a node measures how often a node lies on shortest paths between all other nodes in the graph

BC(1) = 0

BC(2) = 1

BC(3) = 9

BC(4) = 0

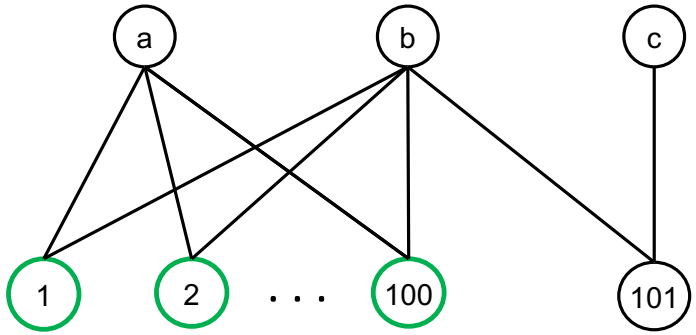


Hypothesis: A value node corresponding to a homograph will have a higher betweenness centrality than a value node with a single meaning

L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1977

BC Speedup (Node Compression)

- BC computation is expensive
  - $O(nm)$ complexity where n is the number of nodes and m is the number of edges in the graph

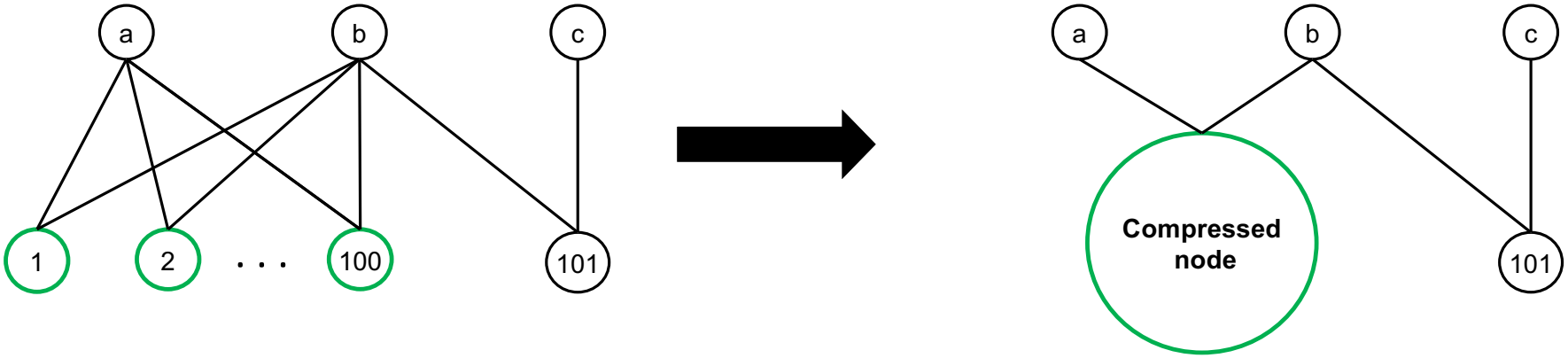


Many nodes always appear in the same set of attributes.

The BC scores of all these nodes is the same!

BC Speedup (Node Compression)

- BC computation is expensive
  - $O(nm)$ complexity where n is the number of nodes and m is the number of edges in the graph

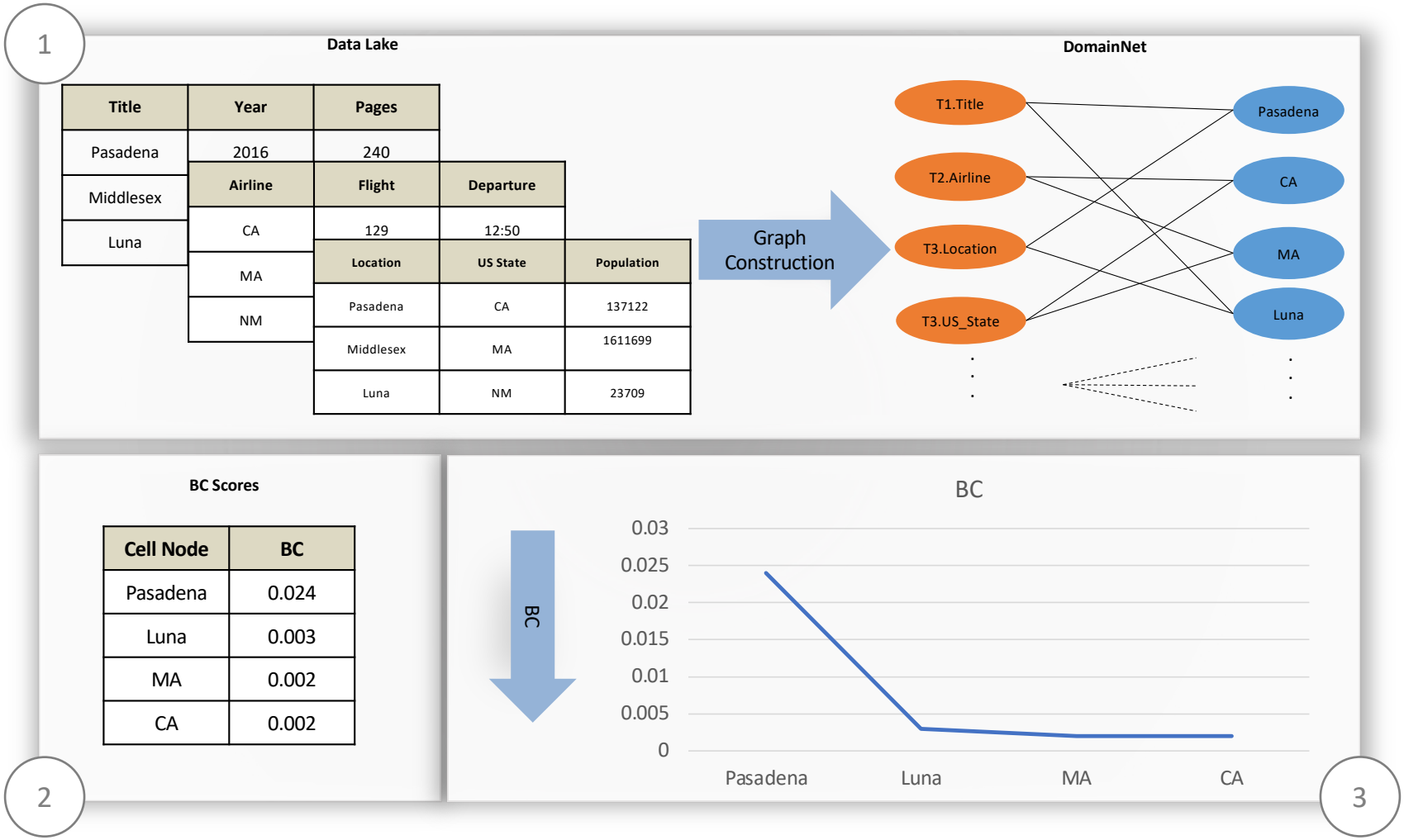


Create compressed nodes to reduce the graph size

Run BC [Sariyüce+13] on the much smaller compressed graph

Sariyüce, A. E., Saule, E., Kaya, K., & Çatalyürek, Ü. V. (2013, May). Shattering and compressing networks for betweenness centrality. In *Proceedings of the 2013 SIAM International Conference on Data Mining* (pp. 686-694). Society for Industrial and Applied Mathematics.

DomainNet Overview



Grouping Attributes of a Homograph by their Meaning

- Problem 1: Given a homograph how many meanings does it have in the data lake?
  - E.g., The data value Cuba has 2 meanings
- Problem 2: Given a homograph group the set of its attributes by their meaning.
  - E.g., Attributes {country1, country2} have one meaning and attributes {title1, title2} have another meaning

country1	capital1
Cuba	Havana
Barbados	Bridgetown
⋮	⋮

country2	capital2
Cuba	Havana
Bahamas	Nassau
⋮	⋮

title1	genre1
Cuba	Havana
Cobra	Nassau
⋮	⋮

title2	genre2
Cuba	Adventure
The Sixth Sense	Drama
⋮	⋮

We need a similarity measure between attributes {country1, country2, title1, title2}

Datasets

	Dataset	#Tables	#Attributes	#Values	#Homographs
Synthetic Benchmark	SB	13	39	17,633	55
Table Union Search	TUS	1,327	9,859	190,399	26,035
TUS Injected	TUS-I	1,253	5,020	163,860	0-500

F. Nargesian, E. Zhu, K. Q. Pu, and R. J. Miller. Table union search on open data. PVLDB, 11(7):813–825, 2018.

Datasets (Synthetic Benchmark)

	Dataset	#Tables	#Attributes	#Values	#Homographs
Synthetic Benchmark	SB	13	39	17,633	55
Table Union Search	TUS	1,327	9,859	190,399	26,035
TUS Injected	TUS-I	1,253	5,020	163,860	0-500

Synthetic Benchmark (SB)

- Made using a data creator that specifies data sources
- Some homographs: *Jaguar* (car or animal), *Lincoln* (car or city or person name)

Datasets (Table Union Search Benchmark)

	Dataset	#Tables	#Attributes	#Values	#Homographs
Synthetic Benchmark	SB	13	39	17,633	55
Table Union Search	TUS	1,327	9,859	190,399	26,035
TUS Injected	TUS-I	1,253	5,020	163,860	0-500

Table Union Search (TUS) benchmark

- Maps each column to a set of other columns that it is unionable with [Nargesian+ 2018]
- We repurpose the benchmark to *derive* a ground truth for homographs

Assumption: A data value is a homograph if it appears in at least two different columns that are not unionable

F. Nargesian, E. Zhu, K. Q. Pu, and R. J. Miller. Table union search on open data. PVLDB, 11(7):813–825, 2018.

Datasets (Table Union Search Benchmark)

	Dataset	#Tables	#Attributes	#Values	#Homographs
Synthetic Benchmark	SB	13	39	17,633	55
Table Union Search	TUS	1,327	9,859	190,399	26,035
TUS Injected	TUS-I	1,253	5,020	163,860	0-500

Table Union Search (TUS) benchmark

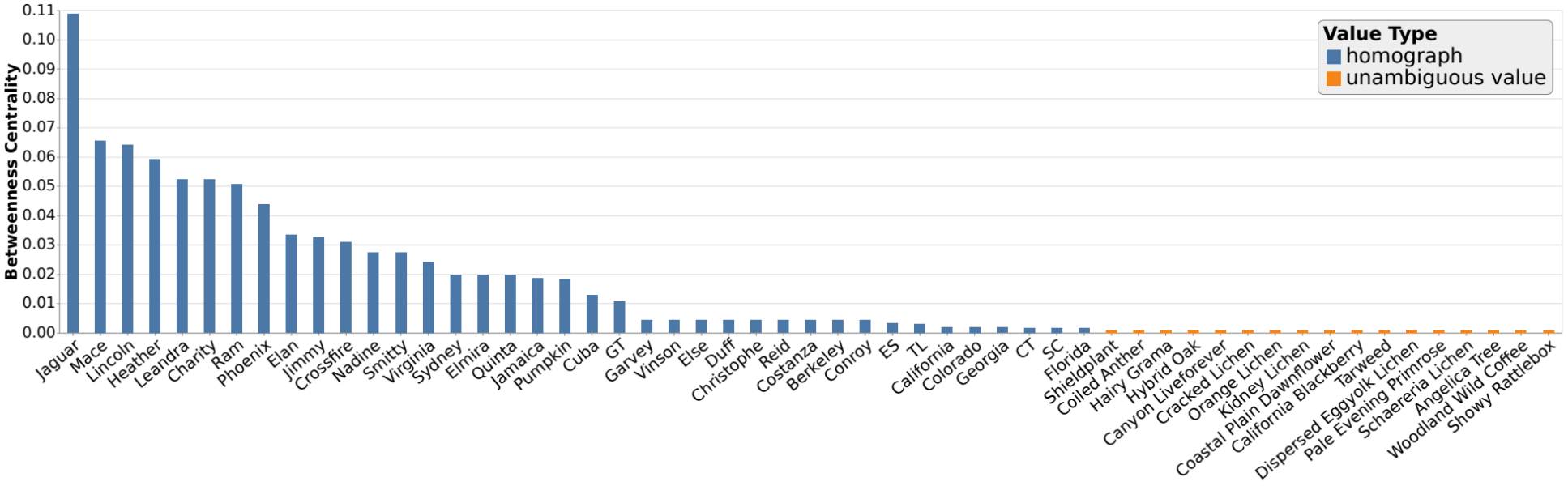
- Maps each column to a set of other columns that it is unionable with [Nargesian+ 2018]
- We repurpose the benchmark to *derive* a ground truth for homographs

Assumption: A data value is a homograph if it appears in at least two different columns that are not unionable

F. Nargesian, E. Zhu, K. Q. Pu, and R. J. Miller. Table union search on open data. PVLDB, 11(7):813–825, 2018.

Experiments (Synthetic Benchmark (SB))

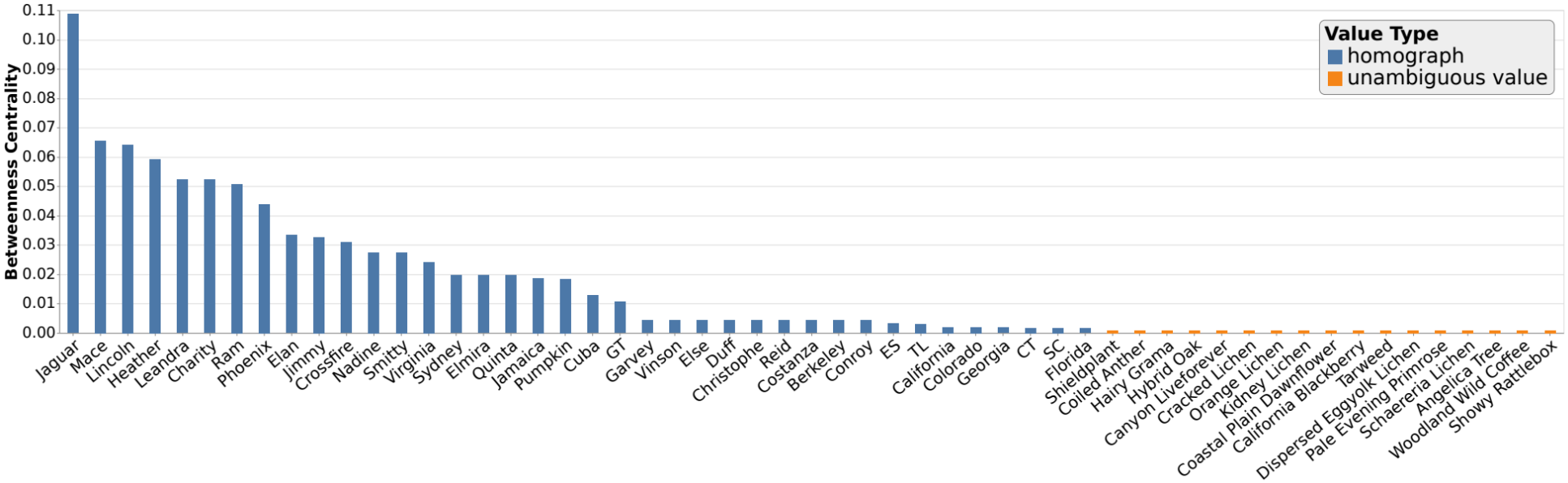
- Rank data values by their BC scores in descending order
- Evaluate at top-55
  - 55 of the 17,633 values are homographs in the SB



Experiments (Synthetic Benchmark (SB))

- Rank data values by their BC scores in descending order
- Evaluate at top-55
  - 55 of the 17,633 values are homographs in the SB

DomainNet identifies 38/55≈**69**% of the homographs

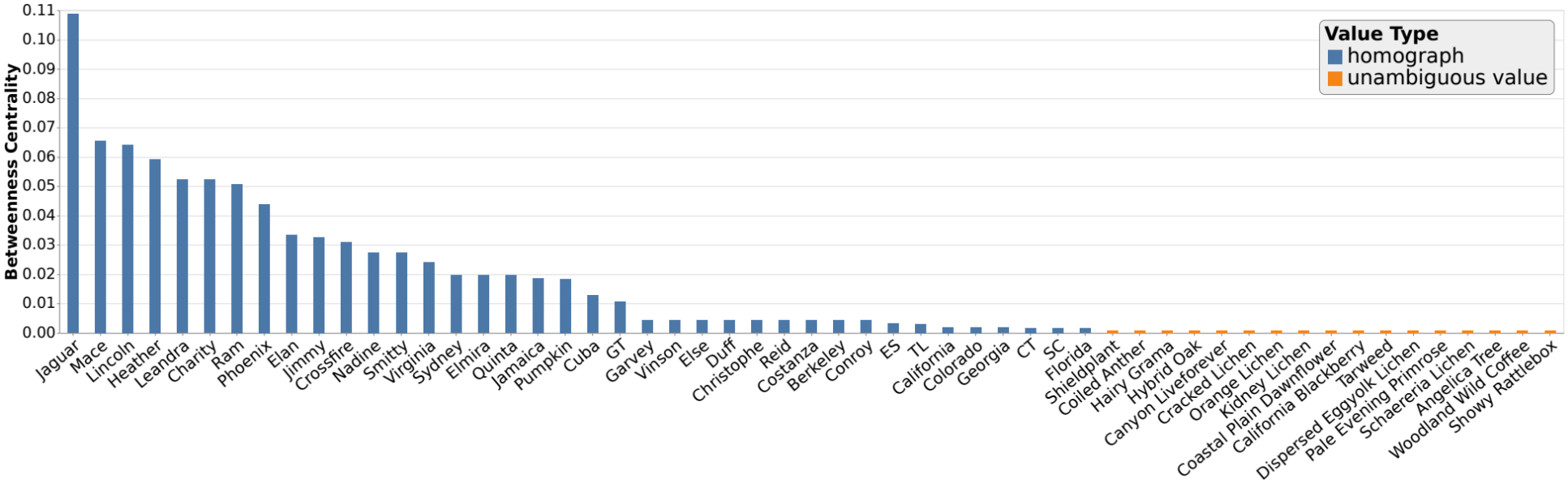


Experiments (Synthetic Benchmark (SB))

- Rank data values by their BC scores in descending order
- Evaluate at top-55
  - 55 of the 17,633 values are homographs in the SB

DomainNet identifies 38/55≈**69**% of the homographs

D^4 identifies 21/55≈**38**% of the homographs



Experiments (Synthetic Benchmark (SB))

Where are the remaining 17 homographs?

Experiments (Synthetic Benchmark (SB))

Where are the remaining 17 homographs?

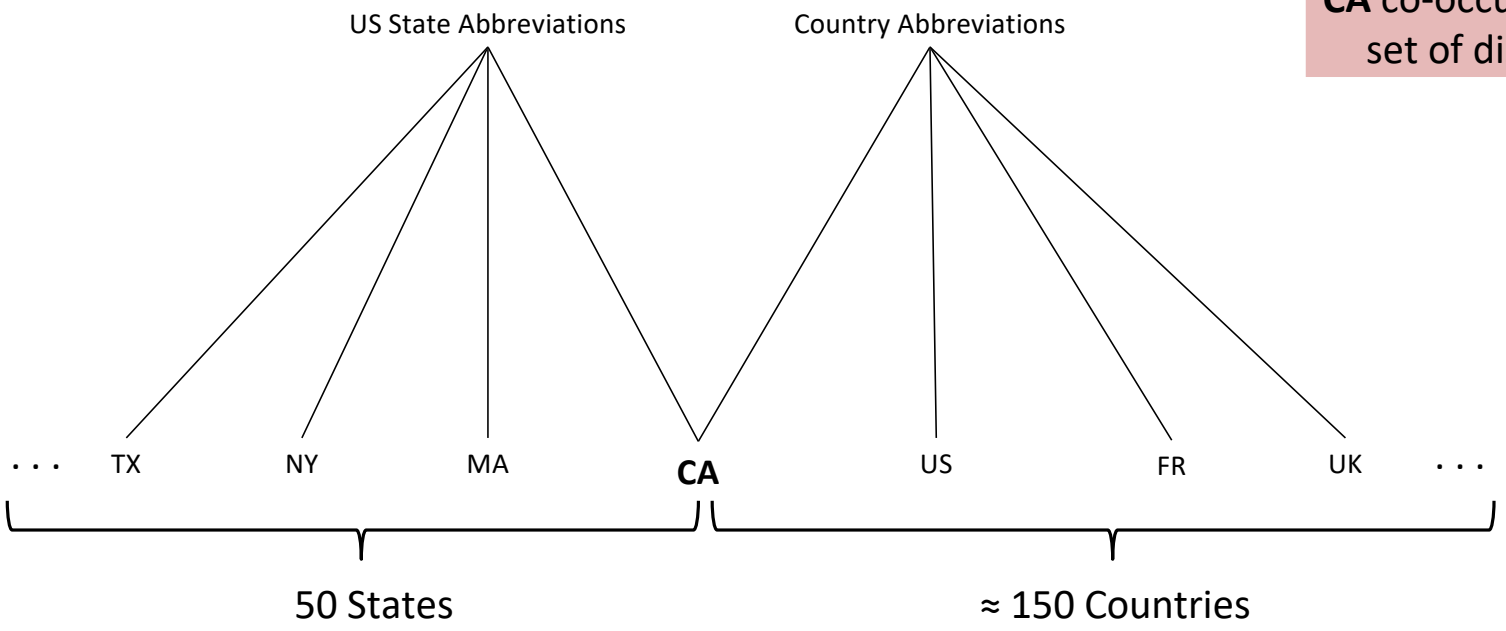
- They correspond to country state name abbreviations (e.g. CA stands for Canada or California)
- They co-occur in a column with a small set of distinct values so fewer shortest paths pass through them

Experiments (Synthetic Benchmark (SB))

Where are the remaining 17 homographs?

- They correspond to country state name abbreviations (e.g. CA stands for Canada or California)
- They co-occur in a column with a small set of distinct values so fewer shortest paths pass through them

CA co-occurs with a small set of distinct values

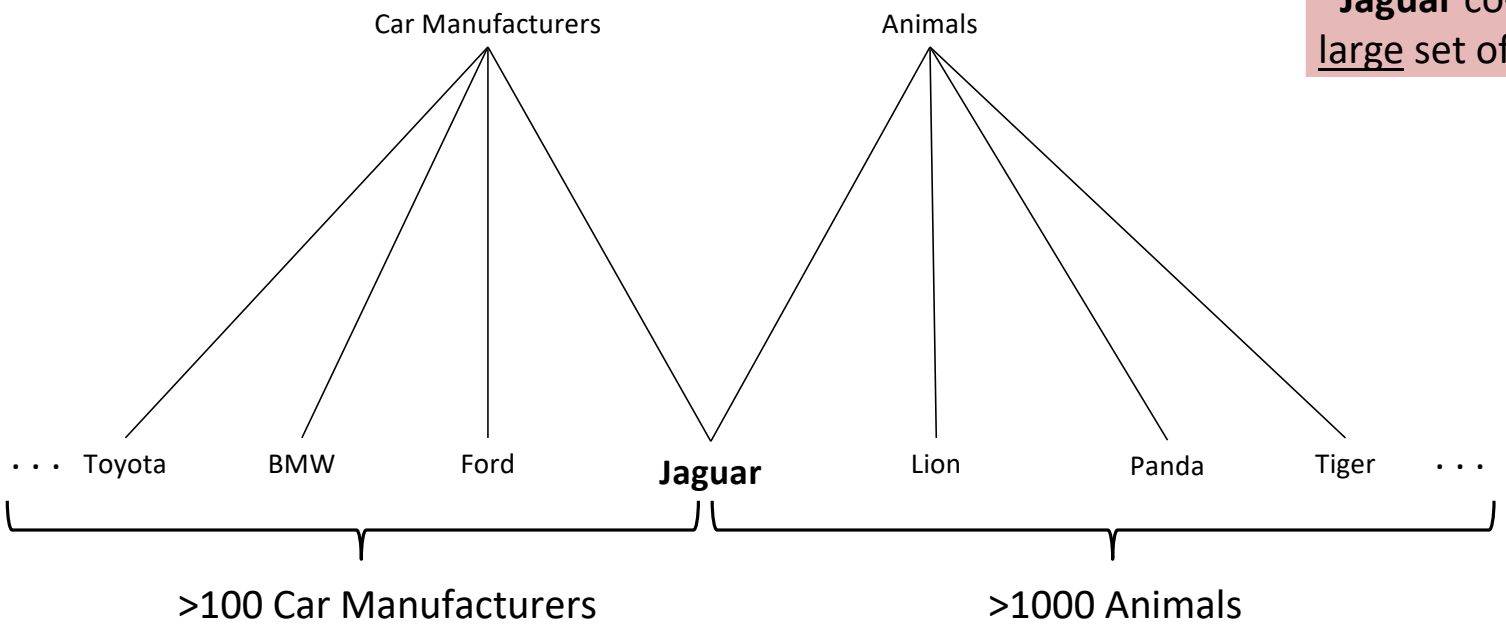


Experiments (Synthetic Benchmark (SB))

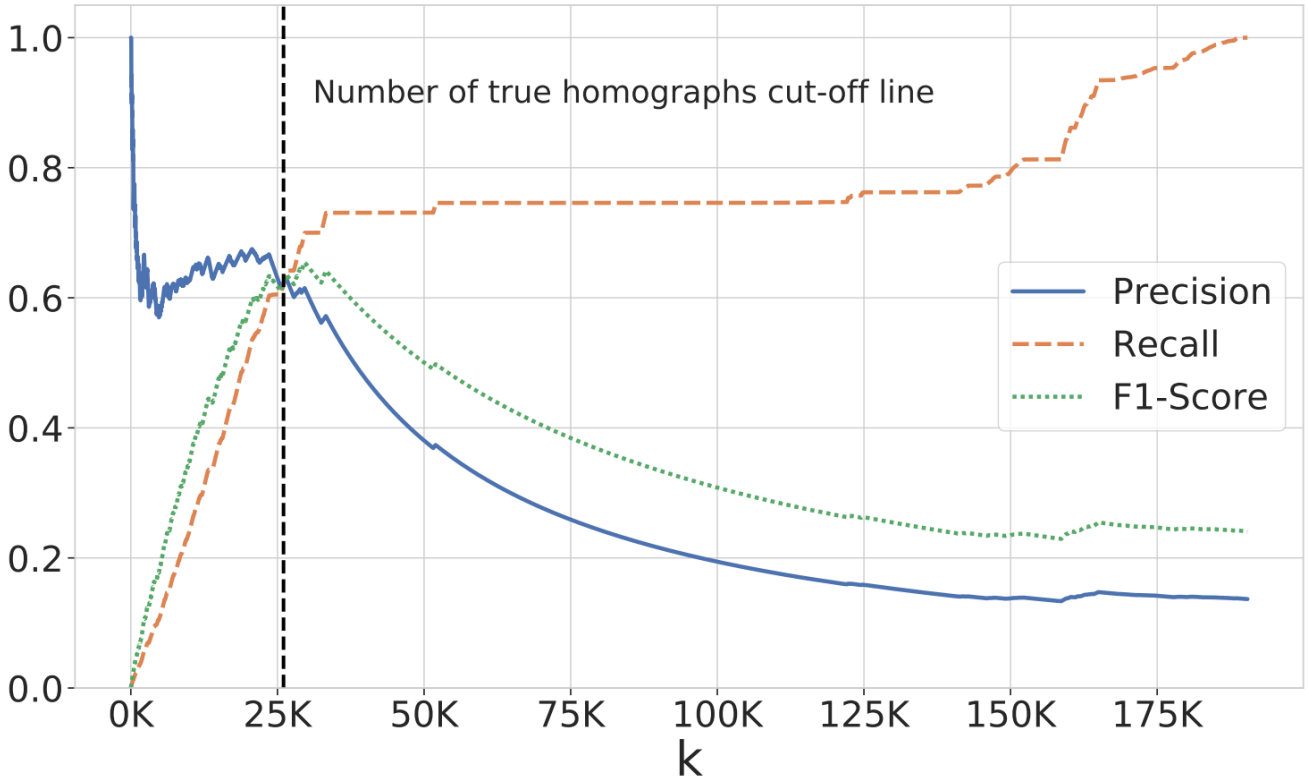
Where are the remaining 17 homographs?

- They correspond to country state name abbreviations (e.g. CA stands for Canada or California)
- They co-occur in a column with a small set of distinct values so fewer shortest paths pass through them

Jaguar co-occurs with a large set of distinct values



Experiments (Table Union Search (TUS) Benchmark)

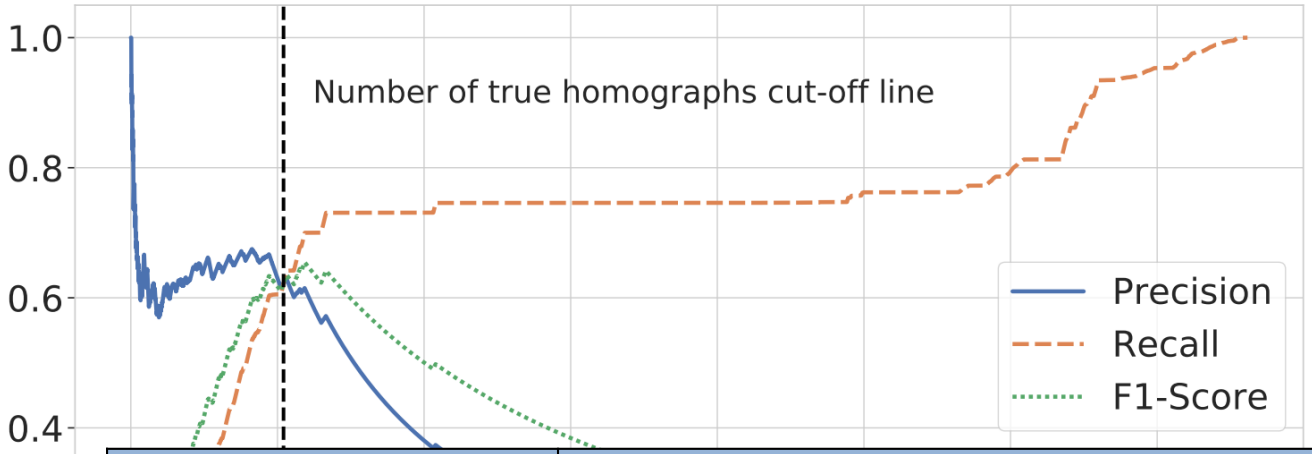


26035 homographs in dataset

Precision at k=200: **0.89**

F1-score at k=26035: **0.622**

Experiments (Example Homographs in TUS Benchmark)



26035 homographs in dataset

Precision at k=200: **0.89**

F1-score at k=26035: **0.622**

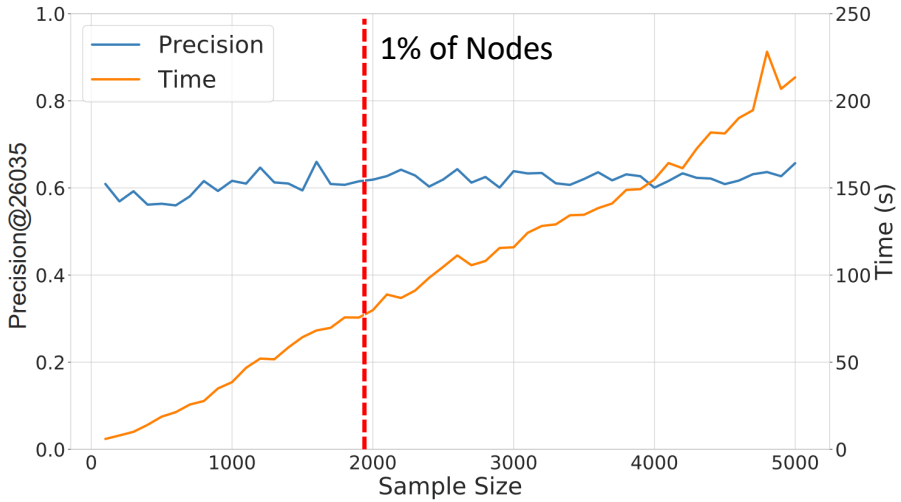
Data Value	Explanation
Music Faculty	University Department *or* Name of a bus stop
.	NULL in many contexts
50	Street number *or* Identifier *or* Quantity etc.
Manitoba Hydro	Electric company *or* Erroneously placed in a street name column

Experiments (Homograph Number of Meanings)

- Given a homograph, can achieve 98% accuracy on determining the number of meanings
  - Experiments vary the parameters to the clustering algorithm (DBSCAN) and show how to set them to achieve this accuracy
- Given a homograph, can achieve 99% accuracy on assigning attributes to a meaning

These results on our Synthetic Benchmark

Experiments (Scalability)

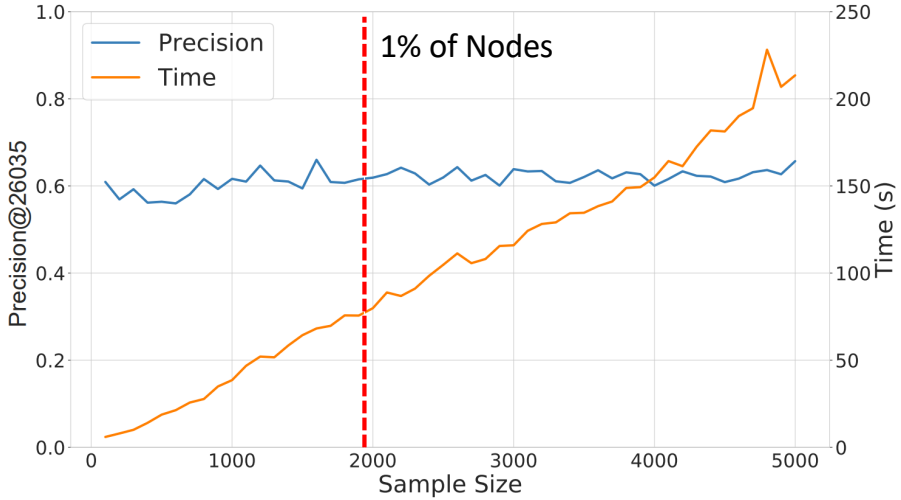


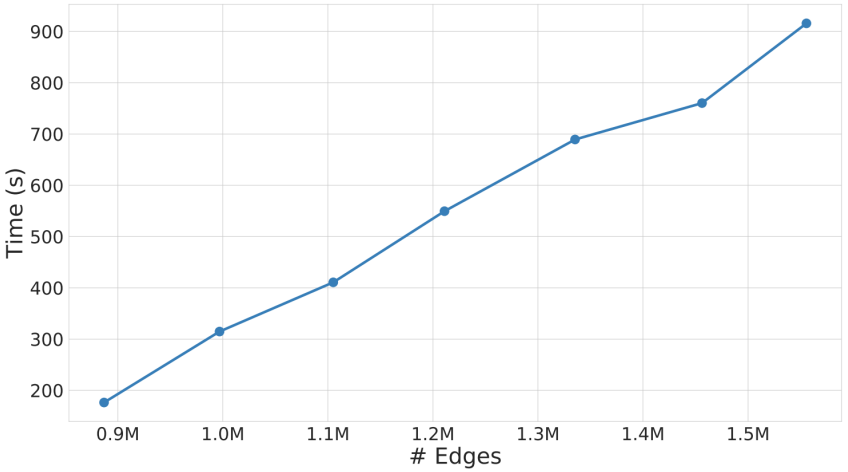
- Exact BC is expensive
- Approximate BC [Geisberger+ 2008]

Sampling more than 1% of the nodes in the TUS benchmark does not significantly change the BC rankings

R. Geisberger, P. Sanders, and D. Schultes. Better approximation of betweenness centrality. In ALENEX, pages 90–100. SIAM, 2008.
Used the NYC Education dataset that is also used by D4 to vary the number of edges. https://zenodo.org/record/3647642

Experiments (Scalability)





- Exact BC is expensive
- Approximate BC [Geisberger+ 2008]

Sampling more than 1% of the nodes in the TUS benchmark does not significantly change the BC rankings

Approximate-BC time complexity is $O(sm)$

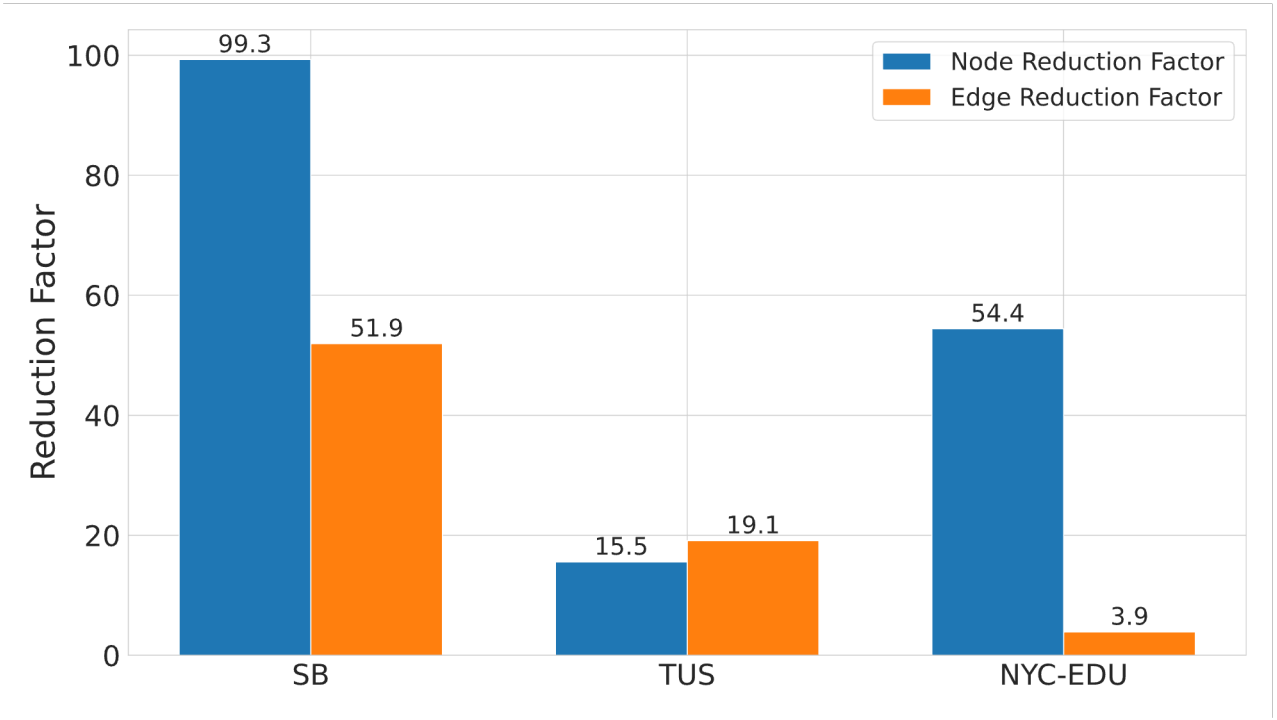
s is the number of nodes sampled

m is the number of edges in the graph

Approximating BC is much faster but still effective

R. Geisberger, P. Sanders, and D. Schultes. Better approximation of betweenness centrality. In ALENEX, pages 90–100. SIAM, 2008.
Used the NYC Education dataset that is also used by D4 to vary the number of edges. https://zenodo.org/record/3647642

Experiments (Scalability - Node Compression)



Dataset	Exact BC runtime without compression	Exact BC runtime with compression
SB	5.52 sec	0.005 sec
TUS	150 min	38 sec
NYC-EDU	>5 hours	162 sec

Note: NYC-EDU has almost 1.5M values, TUS has less than 200K, SB has less than 20K

Experiments (Impact on Domain Detection D^4)

Existence of homographs can negatively impact the performance of existing semantic type detection and domain discovery algorithms such as D^4

In data lake with 1.5 Million values, as few as 150 homographs cut accuracy in half

By **identifying and removing** homographs semantic type detection and domain discovery algorithms like D^4's can be improved

DomainNet Summary

DomainNet: unsupervised technique to identify if a data value is a homograph or not

- Examines data value co-occurrence
- Uses network-centrality measures on a bipartite graph representation of the data lake
- Given a homograph identify its number of meanings and group its attributes by their meaning

Data Lakes do not follow "***unique name assumption***"

different names always refer to different entities in the world

Open Questions & Future Work

- *Can we taxonomize homographs in data lakes ?*
- *Can we resolve synonyms and homographs ?*
- *Can we perform entity resolution in data lakes ?*

Lessons from DomainNet

- This approach is tailored to data lakes with large numbers of heterogeneous tables
- What else can the network of values, attributes, tuples and tables help us understand about how a set of tables in a data lake might best be integrated?
- Embedding approaches for tables also turn tables into graphs
  - Consider pairs of tables [Cappuzzo+20, Li+21]
  - Consider (small) web tables and retrieval/completion rather than integration tasks [Deng+19, Gunther21]

R. Cappuzzo, P. Papotti, S. Thirumuruganathan:Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks. SIGMOD Conference 2020: 1335-1349
Y Li, J. Li, Y. Suhara, J. Wang, W. Hirota, W.-C.Tan: Deep Entity Matching: Challenges and Opportunities. ACM J. Data Inf. Qual. 13(1): 1:1-1:17 (2021)
L. M. Gunther, M. Thiele, J.Gonsior, W. Lehner. Pre-Trained Web Table Embeddings for Table Discovery. Workshop in Exploiting AI Techniques for Data Management}, 2021
Li Deng, Shuo Zhang, Krisztian Balog: Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval. CoRR abs/1906.00041 (2019)

How Do We Study Data Lakes?

- Some data lakes we've used

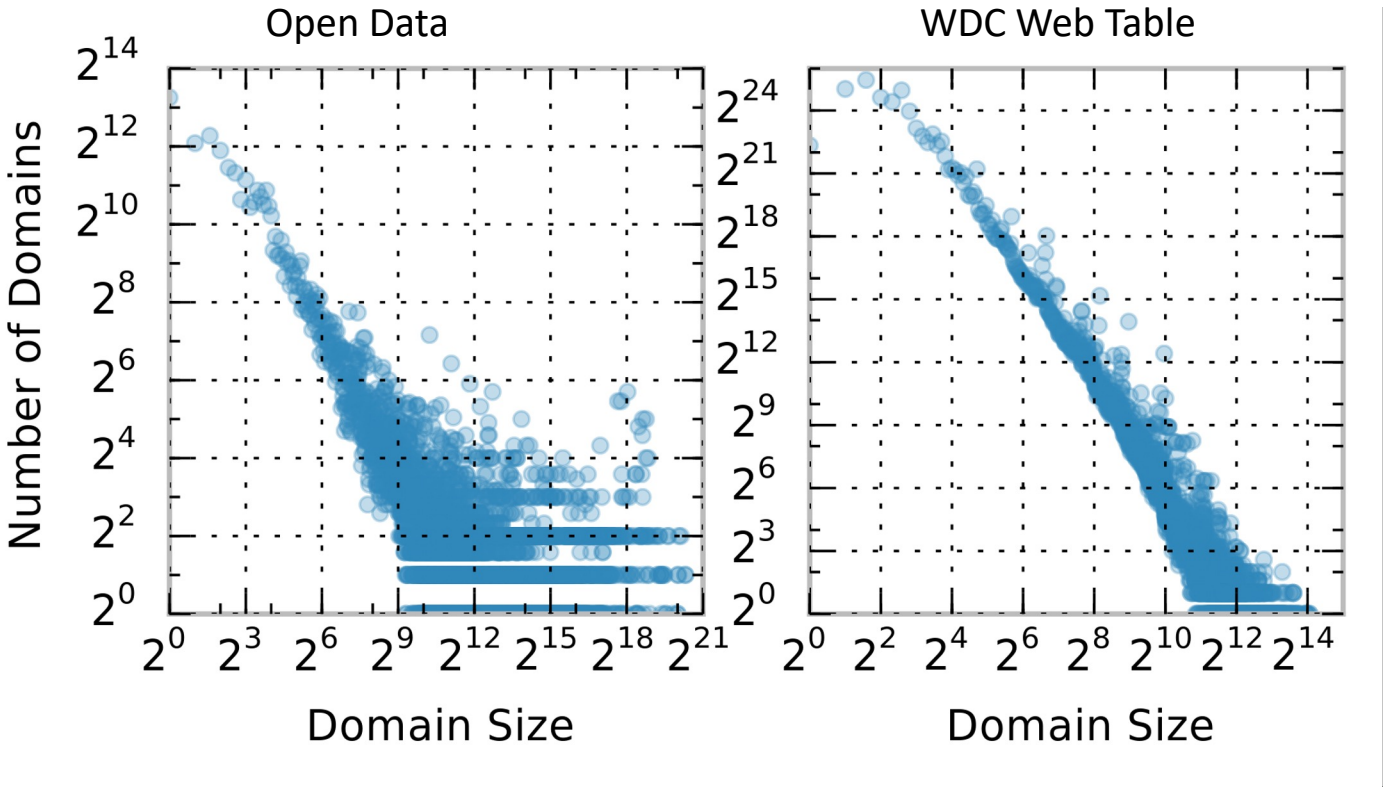
	Avg #Attr	#Attr	MaxSize	AvgSize	AvgStrSize	#UniqVal
OpenData[1,2]	16	3,367,520	22,075,531	465	1,504	609,020,645
WebTable (WDC)[1,2]	5	252,766,759	17,033	10	11	193,071,505
Enterprise[3] - 7%	12	2,032	859,765	4,011		3,902,604

[3] Enterprise lake is a 167 table subset of MIT's 2400 table data warehouse

- Google's dataset search[4] 30M tables (Nov 2020)
- Need to consider representative benchmarks

1. E. Zhu, F. Nargesian, K. Q. Pu, R. J. Miller: LSH Ensemble: Internet-Scale Domain Search. Proc. VLDB Endow. 9(12): 1185-1196 (2016)
2. E. Zhu, K. Q. Pu, F. Nargesian, R. J. Miller: Interactive Navigation of Open Data Linkages. Proc. VLDB Endow. 10(12): 1837-1840 (2017
3. D. Deng, R. C. Fernandez, Z. Abedjan, S. Wang, M. Stonebraker, A. K. Elmagarmid, I. F. Ilyas, S. Madden, M. Ouzzani, N. Tang: The Data Civilizer System. CIDR 2017
4. O. Benjelloun, S. Chen, N. Noy. Google Dataset Search by the Numbers. ISWC (2) 2020: 667-682

Attribute Cardinalities follow power-law



Exploiting the Knowledge of the Data Lake

- Study of Data Lakes as massive graphs
  - What is their underlying topology?
  - Do data lakes graphs exhibit the properties of social networks?
  - Are they naturally small-world or scale-free or …?
- Graph derivations
  - DomainNet - bipartite graph of attributes and values
  - Other graphs can include tuple and table information
- Application of network science to data lakes
  - Community detection, centrality measures, and more
- Data Lakes can challenge and advance network science
  - Can we scale non-parameterized community detection to permit millions of communities?

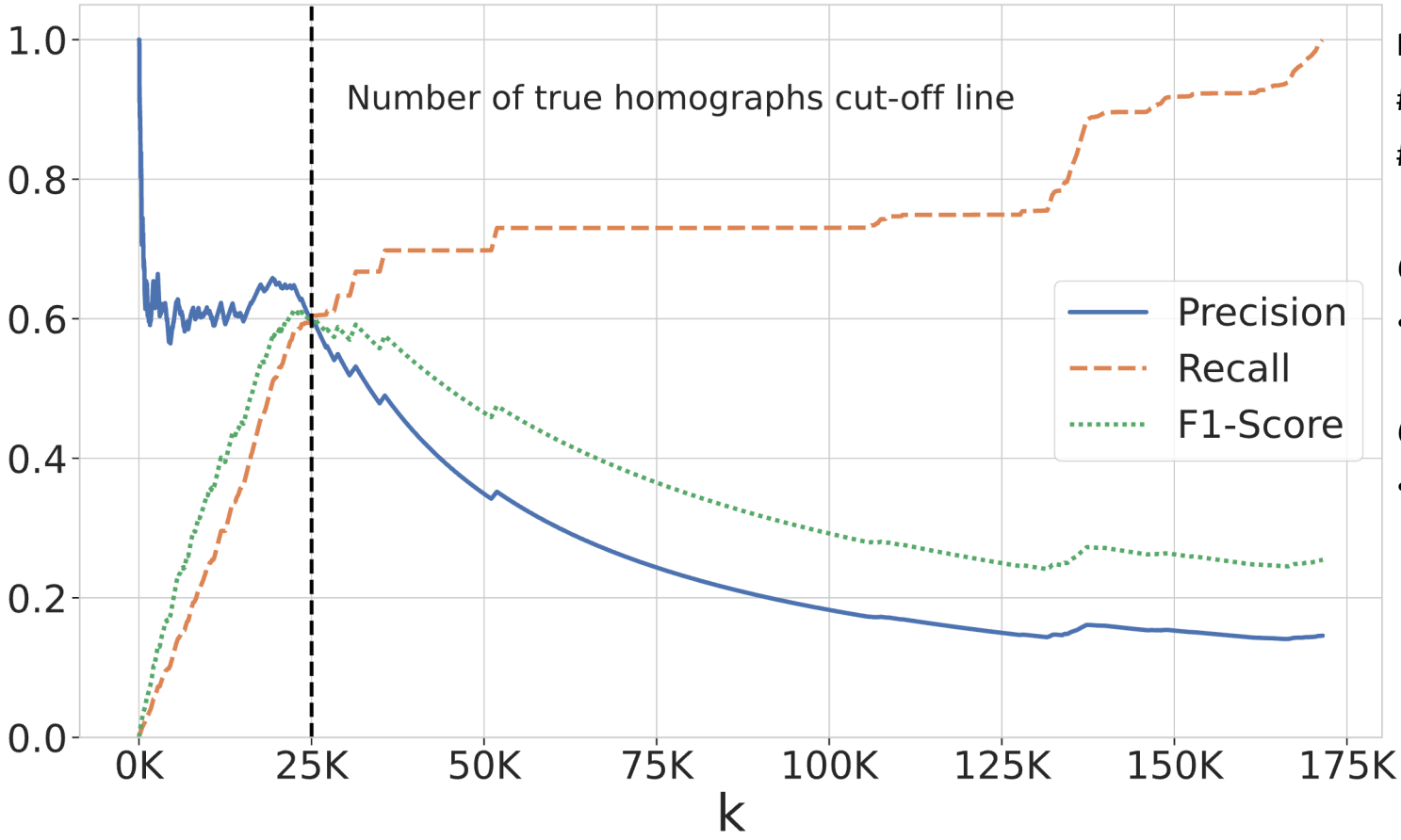
Backup Slides

Synthetic Dataset (13 tables, each table 1000 rows)

1. City, country
2. Country, country code
3. State, state abbreviation
4. Animal name, scientific name, country
5. Animal name, scientific name
6. Plant name, scientific name, family, country
7. Plant name, scientific name, family
8. Company name, full name, country
9. First name, last name, ssn, gender , country
10. Full name, credit card type, credit card number, email address
11. Car manufacturer, car mode, year, country code
12. Grocery item , country
13. Movie title, movie genre, country code

Homograph examples: Sydney (city or name), Jamaica (city or country), Lincoln (car or city), CA (country or state abbreviation), Pumpkin (grocery product or movie title) etc.

TUS Benchmarks with numerical values removed



Numerical Values: 18884

Homographs: 25009

Identical Values: 146506

@k=200:

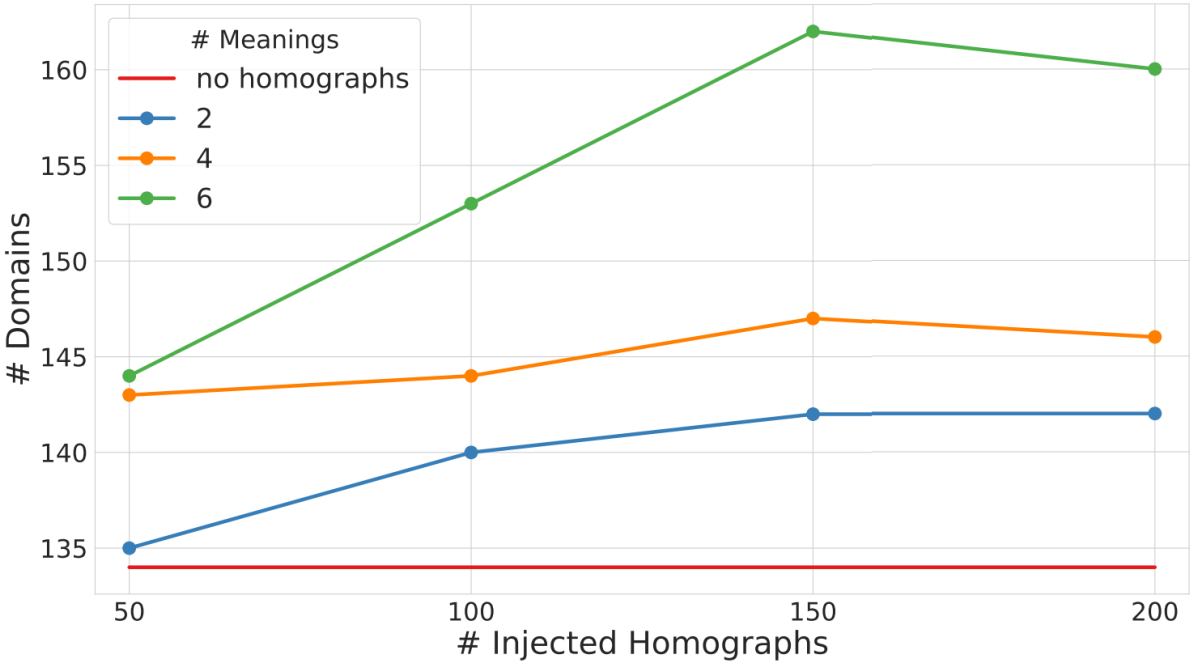
- Precision: 0.875

@k=25009:

- F1-score: 0.594

Experiments (Impact on D^4)

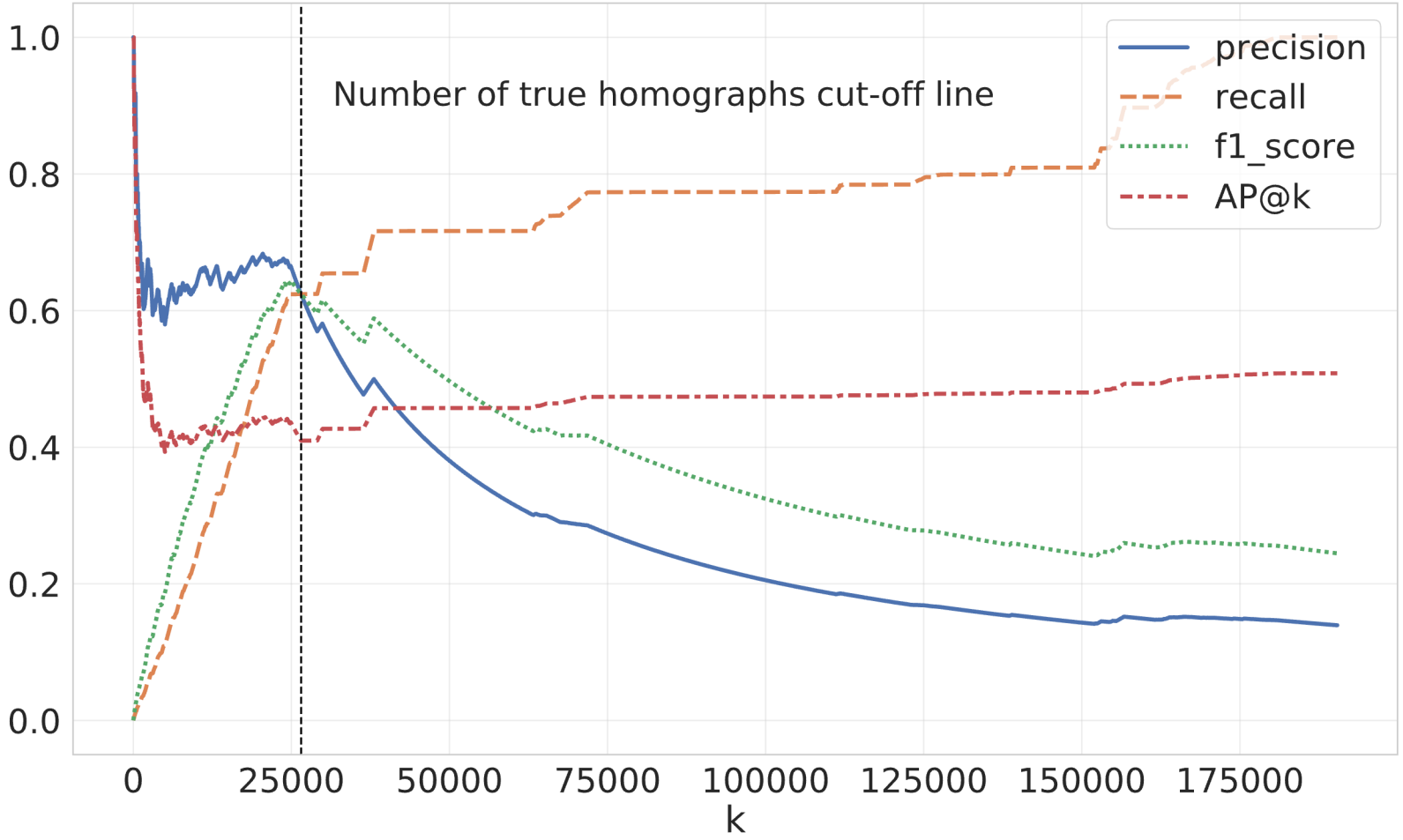
Homograph Injection: Select *n* data values from *n* non-unionable columns and replace them with a new unique value. This value is now a homograph with *n* meanings



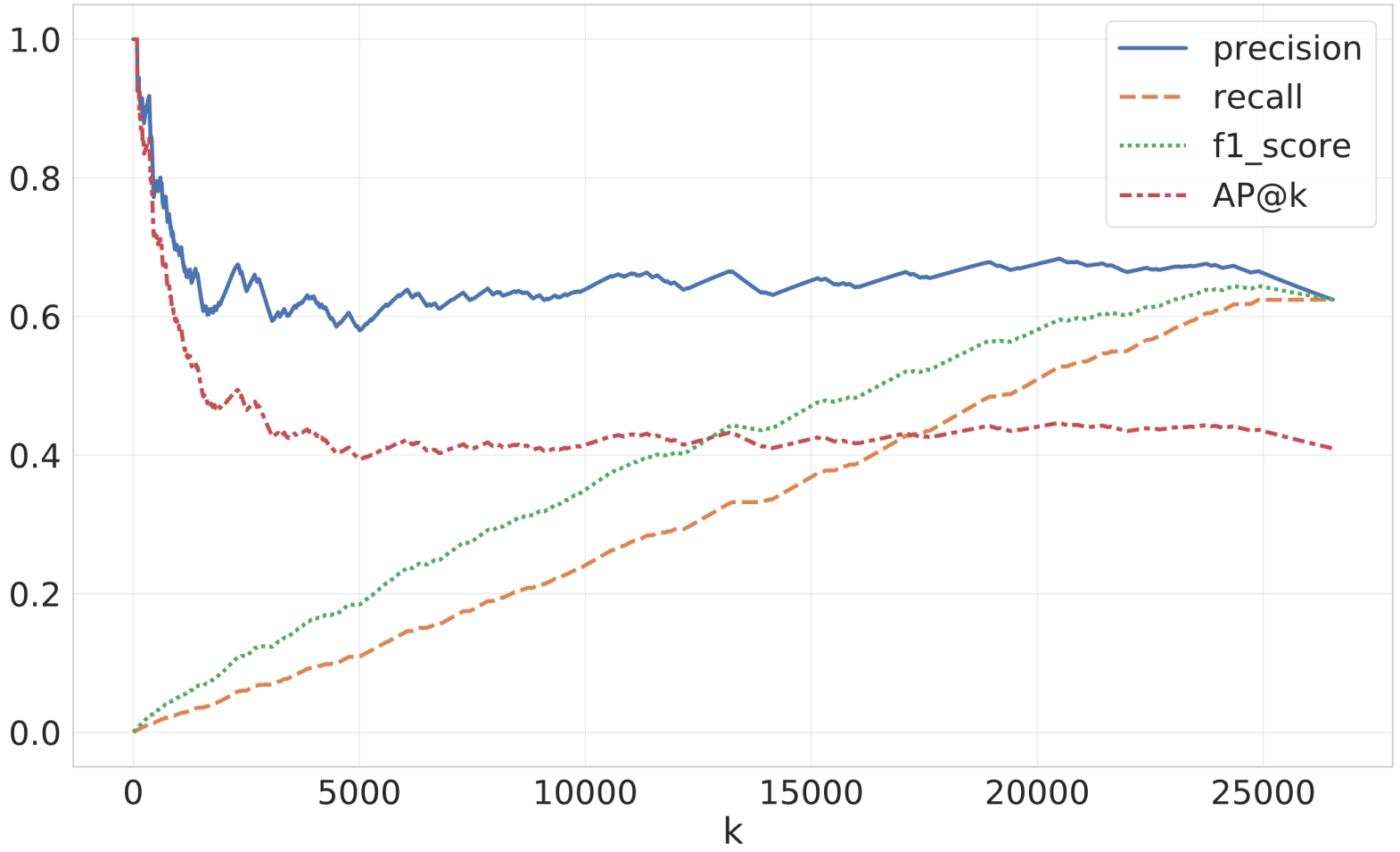
Existence of homographs can negatively impact performance of existing methods such as D^4

By **identifying and removing** homographs D^4's accuracy can be improved

Graph on TUS Benchmark wit AP@k (FULL)



Graph on TUS Benchmark wit AP@k (up to k=26,035)



Variation of source/target nodes with BC

- Modified Networkit package to support BC on a subset of nodes (fast, C++ implementation)
- Synthetic Benchmark: All nodes vs. cell-value-nodes results to the same rankings at top-55

TUS Benchmark

At k= 26035 (# of true homographs), over 5 runs with 5000 sampled nodes

Measure	All Nodes	Cell-Value-Nodes
Precision/Recall/ F1-score	0.622 (σ = 0.008)	0.591 (σ = 0.009)

Rank with the highest F1-score average statistics, over 5 runs with 5000 sampled nodes

Measure	All Nodes	Cell-Value-Nodes
Rank with highest F1-score	27870 (σ = 2617)	22827.6 (σ = 96)
Precision	0.641 (σ = 0.017)	0.673 (σ = 0.002)
Recall	0.672 (σ = 0.052)	0.578 (σ = 0.003)
F1-score	0.654 (σ = 0.022)	0.622 (σ = 0.002)

Using all nodes seems to return slightly better results but with higher standard deviation

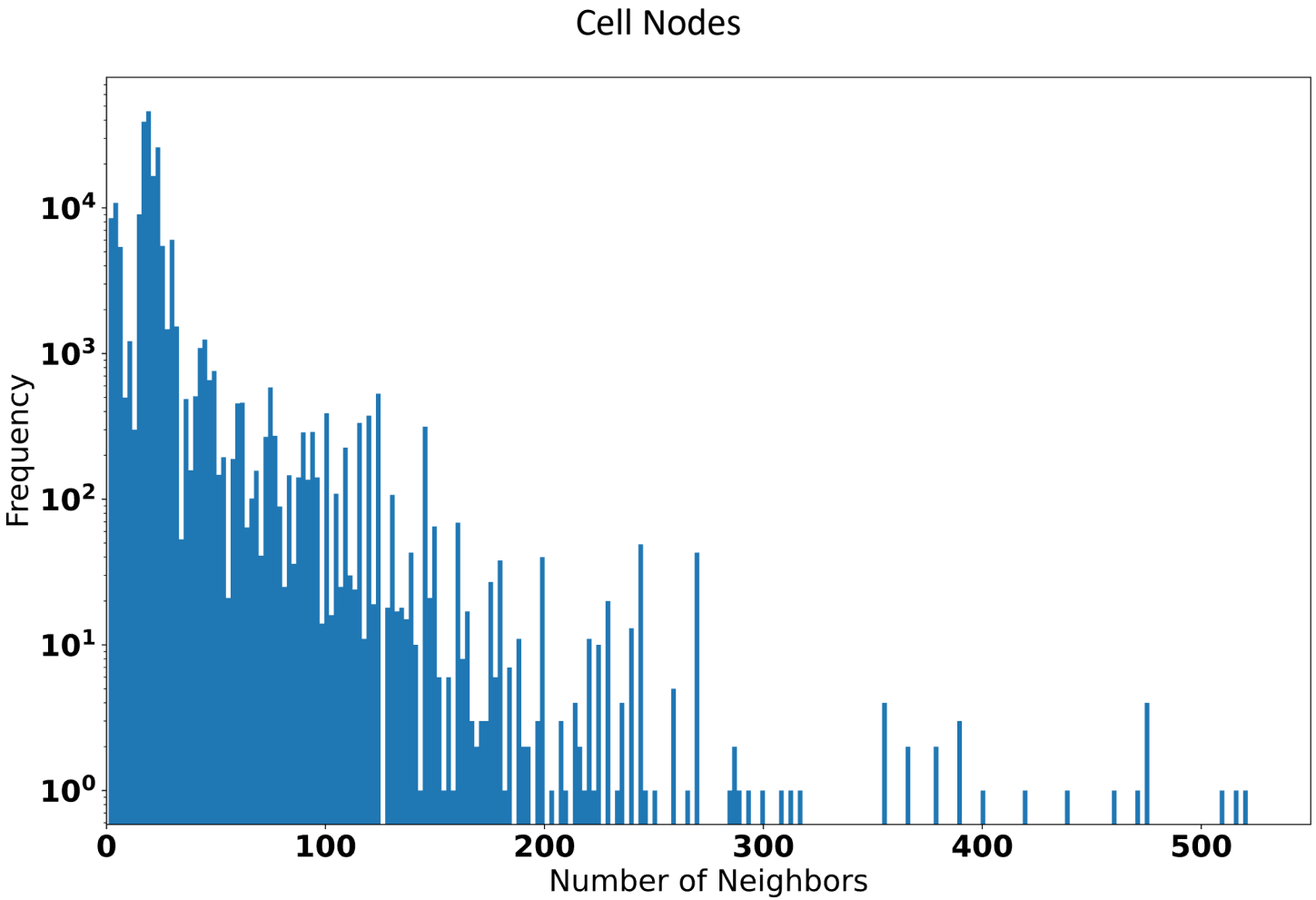
Variation of source/target nodes with BC (All combinations)

Experimenting with all combinations of source/target nodes when approximating BC. All experimental setups (each row) were run 5 times and results are averaged accompanied with their respective sample standard deviation (shown in parentheses)

The cutoff is at k=26035 (# of true homographs) based on the ground truth. @max f1-score refers to the recorded measures at the rank where the maximum f1-score occurs.

source->target	@cutoff f1-score	@max f1-score rank	@max f1-score precision	@max f1-score recall	@max f1-score f1-score
all->all	**0.622** (0.008)	27793.4 (2765)	0.631 (0.023)	0.672 (0.052)	0.650 (0.021)
all->attr	0.622 (0.008)	27693.8 (2959)	0.633 (0.026)	0.671 (0.054)	0.650 (0.021)
all->cell	0.600 (0.006)	23294.6 (40)	0.667 (0.001)	0.597 (0.002)	0.630 (0.001)
attr->all	0.619 (0.003)	29802.0 (6423)	0.614 (0.063)	0.691 (0.084)	0.644 (0.006)
attr->attr	0.619 (0.009)	31894.0 (2889)	0.595 (0.024)	**0.727** (0.040)	**0.653** (0.007)
attr->cell	0.599 (0.000)	23355.0 (133)	0.667 (0.003)	0.599 (0.001)	0.631 (0.000)
cell->all	0.595 (0.007)	22947.2 (106)	**0.668** (0.002)	0.588 (0.004)	0.626 (0.003)
cell->attr	0.595 (0.007)	22946.6 (107)	**0.668** (0.002)	0.588 (0.004)	0.626 (0.003)
cell->cell	0.591 (0.009)	22827.6 (96)	0.667 (0.002)	0.585 (0.003)	0.623 (0.003)

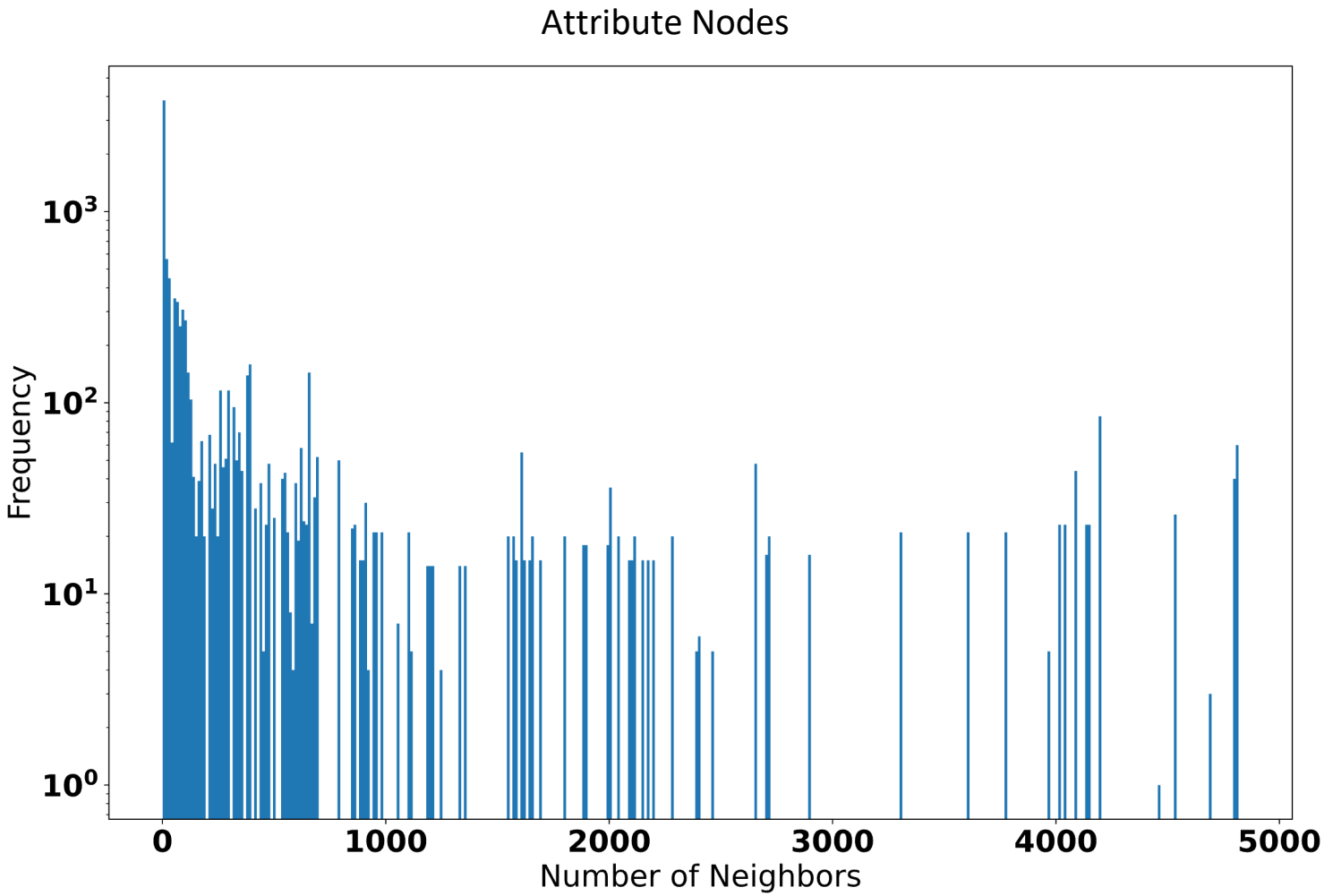
Number of neighbors distribution based on node type



Mean: 21.8

Median: 20

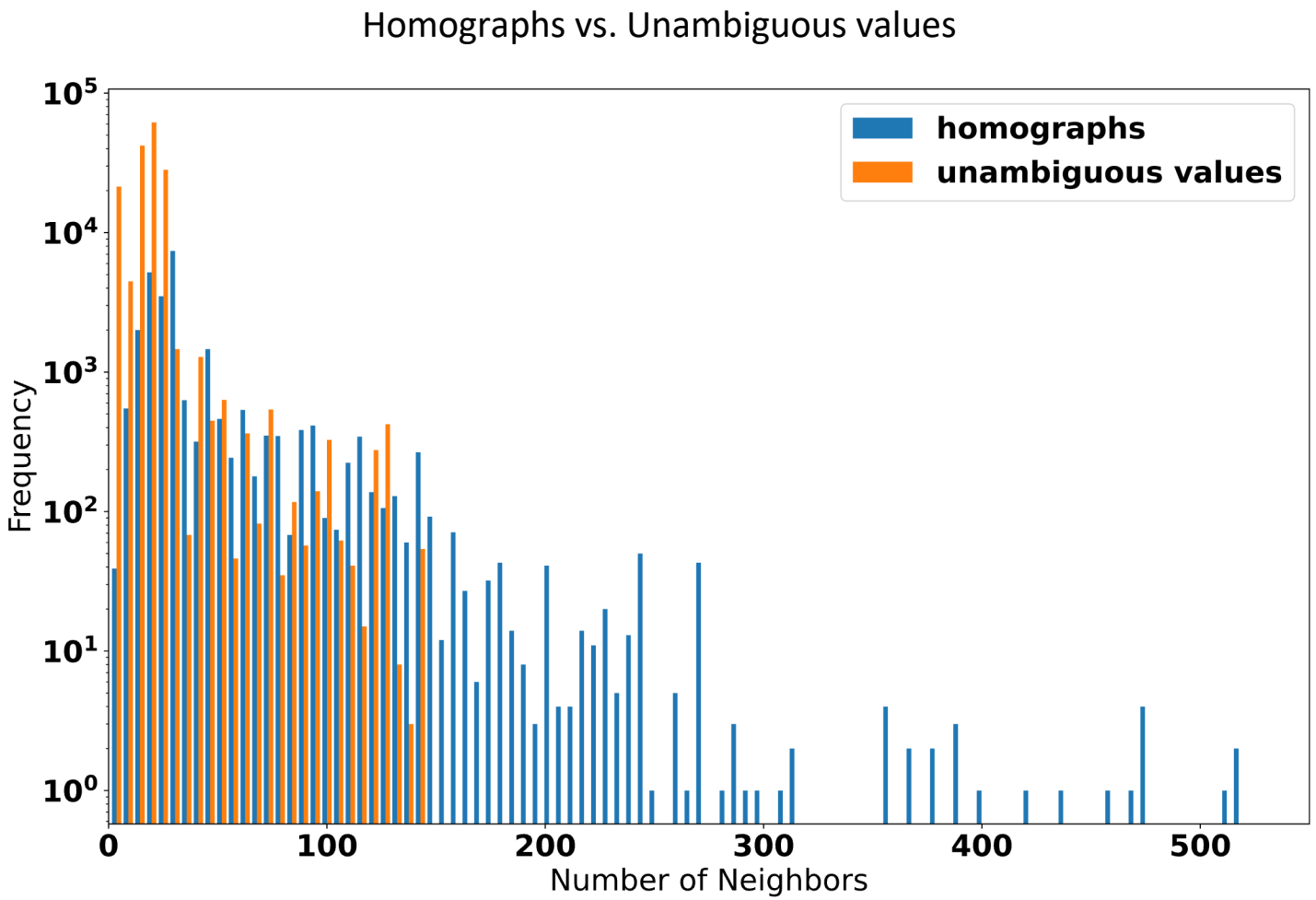
Number of neighbors distribution based on node type



Mean: 421

Median: 52

Number of neighbors distribution based on node type



Homograph Mean: 39.9

Homograph Median: 29

Unambiguous Mean: 18.9

Unambiguous Median: 20

Northeastern University
Khoury College of Computer Sciences

DATALAB @Northeastern

DomainNet: Homograph Detection for Data Lake Disambiguation

Aristotelis Leventidis, Laura Di Rocco, Wolfgang Gatterbauer, Renée J. Miller, Mirek Riedewald

Project Page: https://northeastern-datalab.github.io/table-as-query/
Code: https://github.com/northeastern-datalab/domain_net
Data Lab: https://db.khoury.northeastern.edu/

Data Lakes

Data Lakes are deeply heterogenous where the same data value can have multiple meanings

An Example Data Lake

Col1	Col2	Col3
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

name_1	name_2	name_3
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

An Example Data Lake

Col1	Col2	Col3
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

name_1	name_2	name_3
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

An Example Data Lake

Col1	Col2	Col3
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

name_1	name_2	name_3
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

An Example Data Lake

Col1	Col2	Col3
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

name_1	name_2	name_3
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

An Example Data Lake

Title	Year	Pages
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

name_1	name_2	name_3
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

An Example Data Lake

Title	Year	Pages
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

Location	State	Population
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

An Example Data Lake

Title	Year	Pages
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

Location	State	Population
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

	Airline	Flight	Departure
Air China	CA	129	12:50
Malév Hungarian Airlines	MA	355	16:45
Air Madrid	NM	280	7:35

An Example Data Lake

Homograph: A data value in the data lake with more than one meaning

Title	Year	Pages
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

Location	State	Population
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

Airline	Flight	Departure
CA	129	12:50
MA	355	16:45
NM	280	7:35

DomainNet

Data Lake Disambiguation: Given a data lake containing a collection of tables with possibly missing, incomplete and/or heterogenous table and attribute names. Determine if a data value v that appears in more than one attribute or table has a single meaning or more than one meaning.

- Novel problem: Homograph detection in data lakes
- **DomainNet**: unsupervised technique to identify if a data value is a homograph or not
  - Examines data value co-occurrence
  - Uses network-centrality measures on a bipartite graph representation of the data lake
- First benchmarks (both real and synthetic data) for homograph detection in data lakes

Disambiguation in Literature

Entity Resolution (ER)

- Do two tuples refer to the same real-world entity?
  - E.g., "X. Wang" one or many authors [Bhattacharya+ 2007]
- Resolution is between entities of the same type ☹
  - i.e., cannot resolve *Pasadena* the book versus the city

Semantic Type Detection

- Knowledge-based Techniques
  - Low coverage in data lakes [Nargesian+ 2018]
- Supervised Techniques
  - Sherlock [Hulsebos+ 2019], SATO [Zhang+ 2020]
  - Trained on only 78 types!

- Unsupervised Domain Discovery
  - D^4 [Ota+ 2020]
  - Groups data values to domains
  - We use it as a baseline

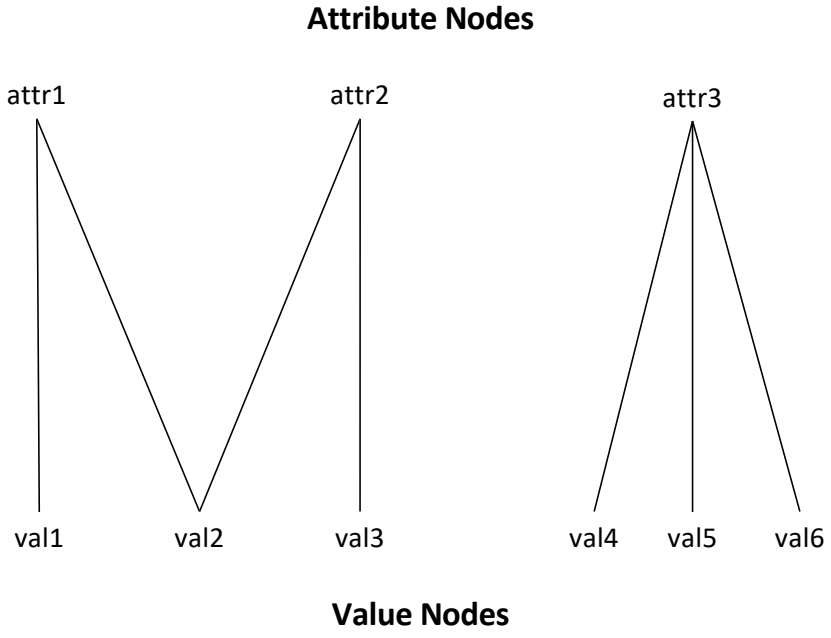
Bhattacharya and L. Getoor. Collective entity resolution in relational data. ACM Trans. Knowl. Discov. Data, 1(1):5, 2007.
F. Nargesian, E. Zhu, K. Q. Pu, and R. J. Miller. Table union search on open data. PVLDB, 11(7):813–825, 2018.
M. Hulsebos, K. Z. Hu, M. A. Bakker, E. Zgraggen, A. Satyanarayan, T. Kraska, Ç. Demiralp, and C. A. Hidalgo. Sherlock: A deep learning approach to semantic data type detection. In SIGKDD, pages 1500–1508. ACM, 2019
D. Zhang, Y. Suhara, J. Li, M. Hulsebos, Ç. Demiralp, and W. Tan. Sato: Contextual semantic type detection in tables. PVLDB, 13(11):1835–1848, 2020
M. Ota, H. Mueller, J. Freire, and D. Srivastava. Data-driven domain discovery for structured datasets. PVLDB, 13(7):953–965, 2020.

DomainNet (A graph Representation)

Intuition: A *homograph* likely co-occurs with a set of values that do not co-occur frequently with each other

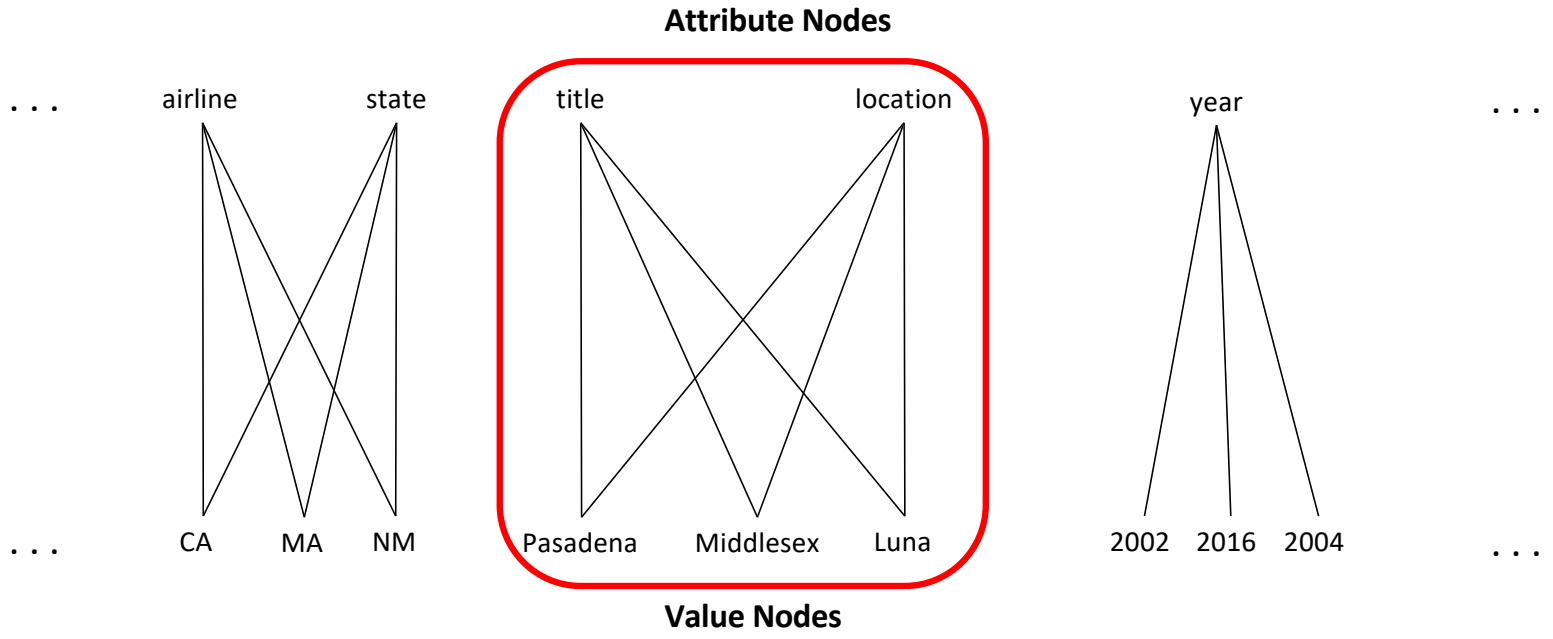
Bipartite graph (Attributes & Values)

- Edge denotes that a value appears with an attribute
- Set semantics for value nodes
- Bag semantics for attribute nodes
- The data value *val2* is found in attributes attr1 and attr2



DomainNet (A graph Representation)

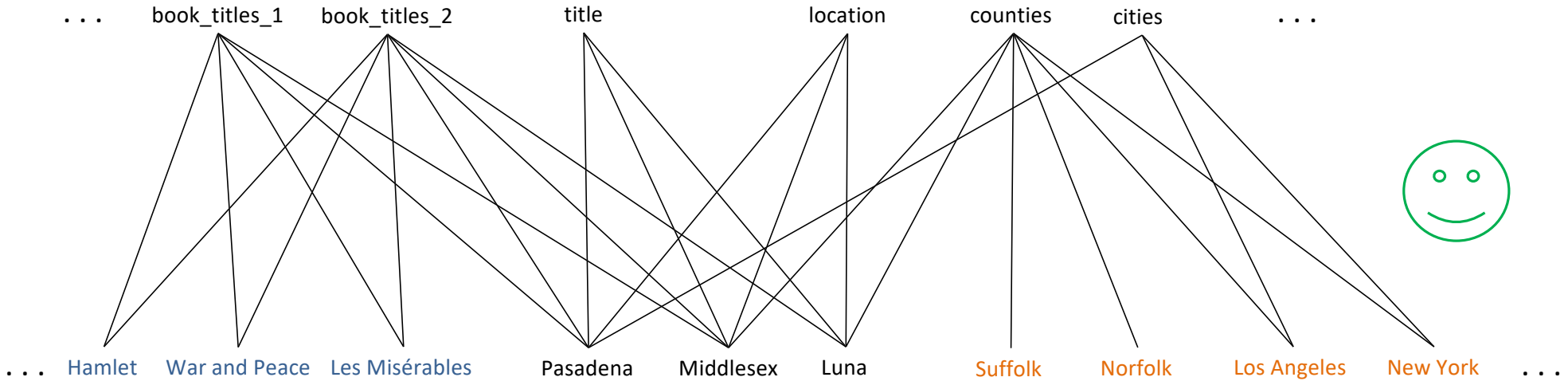
Intuition: A *homograph* likely co-occurs with a set of values that do not co-occur frequently with each other



DomainNet (A graph Representation)

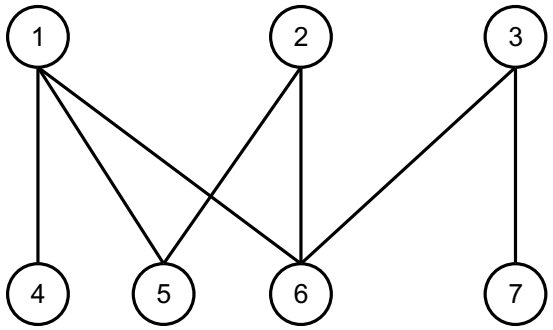
Intuition: A *homograph* likely co-occurs with a set of values that do not co-occur frequently with each other

With enough data the co-occurrence pattern should emerge



Betweenness Centrality (BC)

- Betweenness centrality (BC) [Freeman 1977] of a node measures how often a node lies on paths between all other nodes in the graph
- Numerous applications in telecommunications, social networks, biology etc.

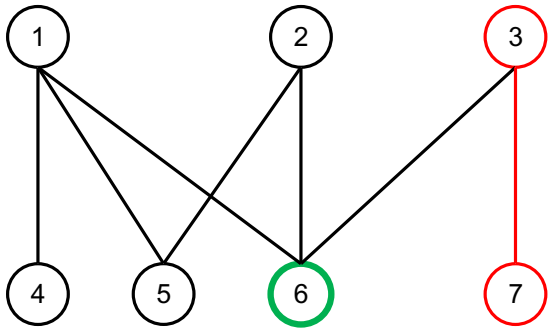


L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1977

Betweenness Centrality (BC)

- Betweenness centrality (BC) [Freeman 1977] of a node measures how often a node lies on paths between all other nodes in the graph
- Numerous applications in telecommunications, social networks, biology etc.

$$\mathrm{BC}(6) = \sum_{v \neq 6,\, w \neq 6} \frac{\sigma_{vw}(6)}{\sigma_{vw}} = 0 + \cdots$$



$$\frac{\sigma_{3,7}(6)}{\sigma_{3,7}} = 0$$

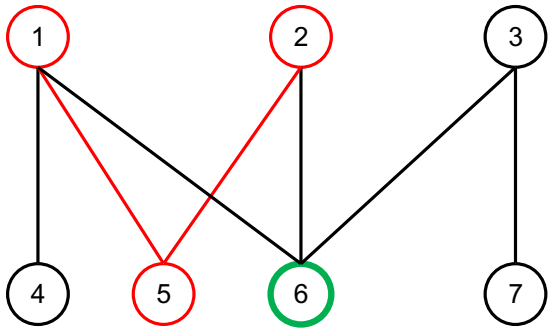
σ_{vw} are the total number of shortest paths from v to w and $\sigma_{vw}(u)$ are the number of shortest paths from v to w that pass through u

L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1977

Betweenness Centrality (BC)

- Betweenness centrality (BC) [Freeman 1977] of a node measures how often a node lies on paths between all other nodes in the graph
- Numerous applications in telecommunications, social networks, biology etc.

$$\mathrm{BC}(6) = \sum_{v \neq 6,\, w \neq 6} \frac{\sigma_{vw}(6)}{\sigma_{vw}} = 0 + \cdots$$



Shortest paths between nodes 1 and 2:

- 1,5,2

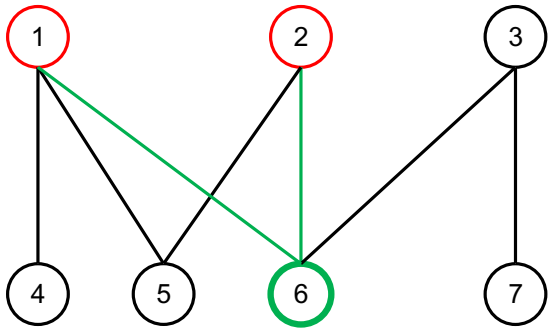
σ_{vw} are the total number of shortest paths from v to w and $\sigma_{vw}(u)$ are the number of shortest paths from v to w that pass through u

L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1977

Betweenness Centrality (BC)

- Betweenness centrality (BC) [Freeman 1977] of a node measures how often a node lies on paths between all other nodes in the graph
- Numerous applications in telecommunications, social networks, biology etc.

$$\mathrm{BC}(6) = \sum_{v \neq 6,\, w \neq 6} \frac{\sigma_{vw}(6)}{\sigma_{vw}} = 0 + \frac{1}{2} + \cdots$$



Shortest paths between nodes 1 and 2:

- 1,5,2
- 1,6,2

$$\frac{\sigma_{1,2}(6)}{\sigma_{1,2}} = \frac{1}{2}$$

σ_{vw} are the total number of shortest paths from v to w and $\sigma_{vw}(u)$ are the number of shortest paths from v to w that pass through u

L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1977

Betweenness Centrality (BC)

- Betweenness centrality (BC) [Freeman 1977] of a node measures how often a node lies on paths between all other nodes in the graph
- Numerous applications in telecommunications, social networks, biology etc.

BC(1) = 6.5

BC(2) = 1.5

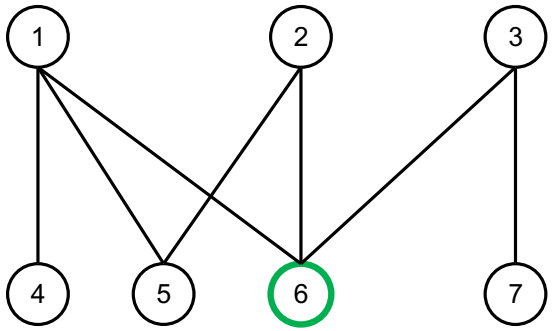
BC(3) = 5

BC(4) = 0

BC(5) = 1

BC(6) = 9

BC(7) = 0



Hypothesis: A value node corresponding to a homograph will have a higher betweenness centrality than a value node with a single meaning

L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1977

Datasets

Dataset	#Tables	#Attributes	#Values	#Homographs
SB	13	39	17,633	55
NYC-EDU	201	3,496	1,469,547	unknown
TUS	1,327	9,859	190,399	26,035
TUS-I	1,253	5,020	163,860	0

Datasets

Dataset	#Tables	#Attributes	#Values	#Homographs
SB	13	39	17,633	55
NYC-EDU	201	3,496	1,469,547	unknown
TUS	1,327	9,859	190,399	26,035
TUS-I	1,253	5,020	163,860	0

Synthetic Benchmark (SB)

- Made using a data creator that specifies data sources
- 55 homographs

Datasets

Dataset	#Tables	#Attributes	#Values	#Homographs
SB	13	39	17,633	55
NYC-EDU	201	3,496	1,469,547	unknown
TUS	1,327	9,859	190,399	26,035
TUS-I	1,253	5,020	163,860	0

Synthetic Benchmark (SB)

- Made using a data creator that specifies data sources
- 55 homographs

New York City Education (NYC-EDU) benchmark

- Large repository of open data used to test the scalability of our method
- Also used by D^4 [Ota+ 2020]

M. Ota, H. Mueller, J. Freire, and D. Srivastava. Data-driven domain discovery for structured datasets. PVLDB, 13(7):953–965, 2020.

Datasets

Dataset	#Tables	#Attributes	#Values	#Homographs
SB	13	39	17,633	55
NYC-EDU	201	3,496	1,469,547	unknown
TUS	1,327	9,859	190,399	26,035
TUS-I	1,253	5,020	163,860	0

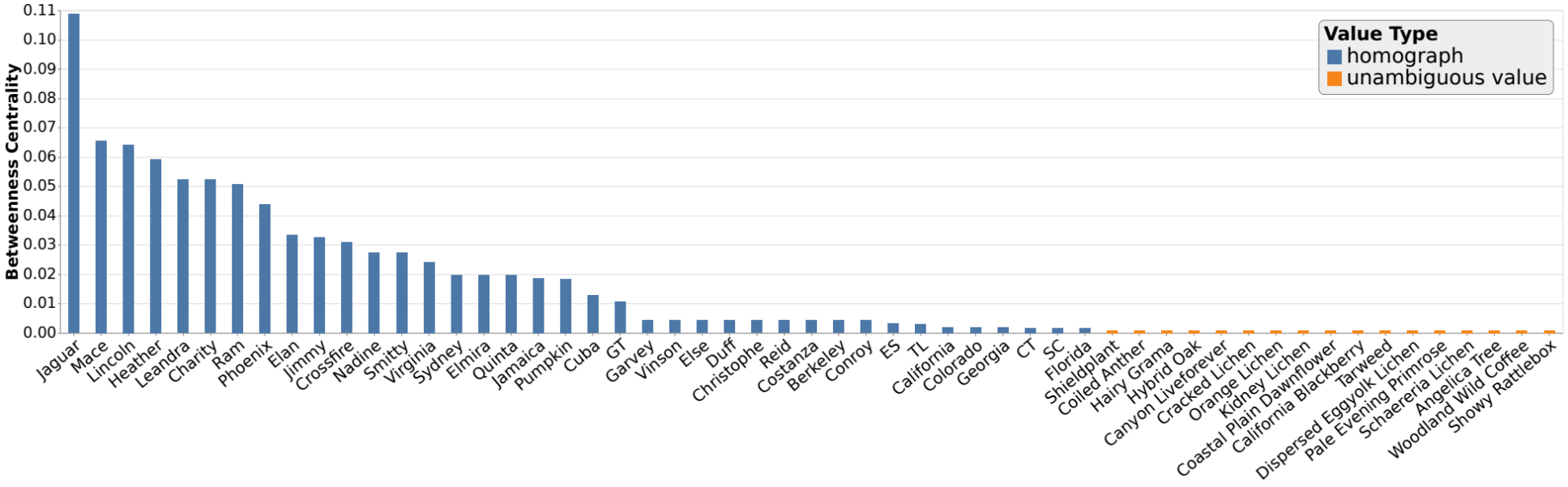
Table Union Search (TUS) benchmark

- Maps each column to a set of other columns that it is unionable with [Nargesian+ 2018]
- We repurpose the benchmark to *derive* a ground truth for homographs
- A data value is a homograph if it appears in at least two different columns that are not unionable
- Derivation of the ground truth is not 100% accurate

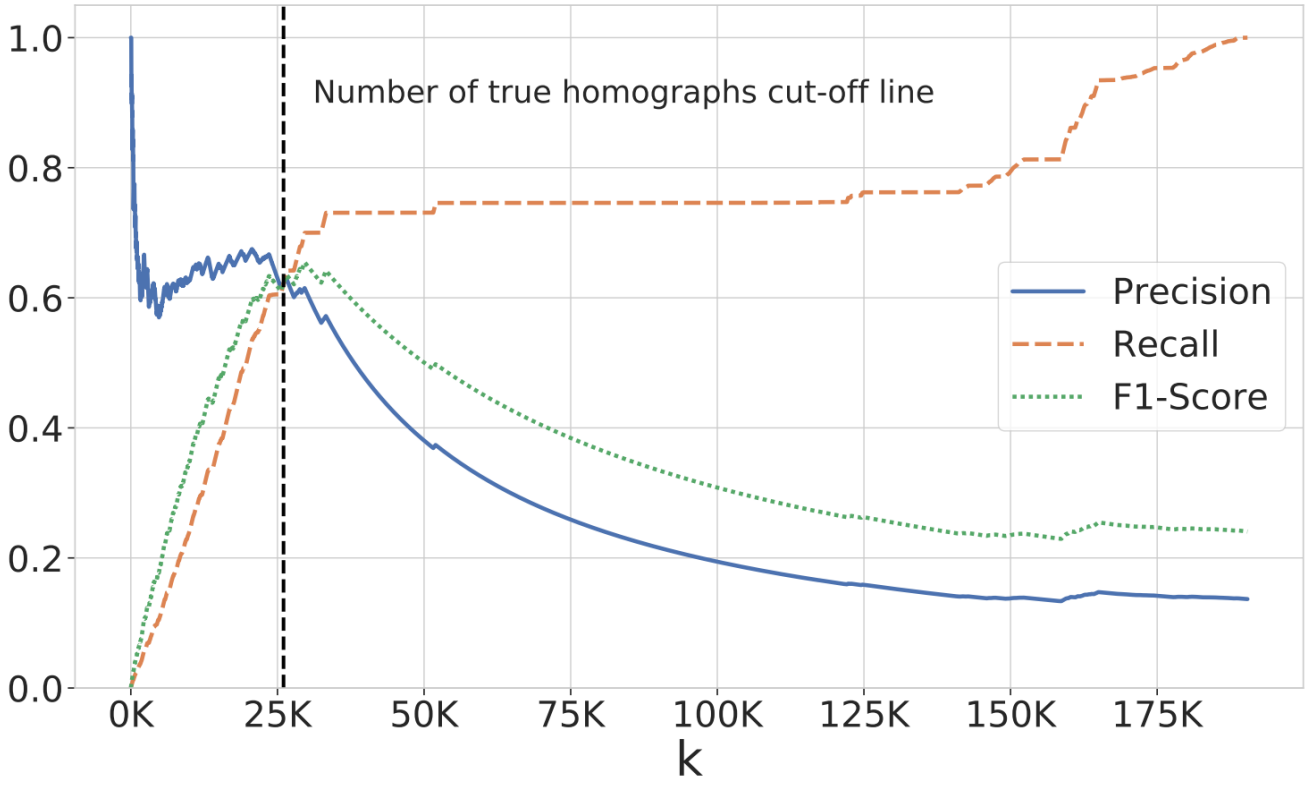
F. Nargesian, E. Zhu, K. Q. Pu, and R. J. Miller. Table union search on open data. PVLDB, 11(7):813–825, 2018.

Experiments (SB)

- Evaluation at top-55. 55 of the 17,633 values are homographs in the SB
- 38/55≈69% are homographs vs. D^4 identifies 21/55≈38% of the homographs
- Where are the remaining 17 homographs?
  - They correspond to country state name abbreviations (e.g. CA stands for Canada or California)
  - They co-occur in a column with a small set of distinct values so fewer shortest paths pass through them



Experiments (TUS)



- 26935 homographs
- Precision at k=200: **0.89**
- Precision/Recall/F1-score at k=26035: **0.622**

<u>Some top values based on BC</u>

Music Faculty

- University Dept. or Location

.

- NULL in many contexts

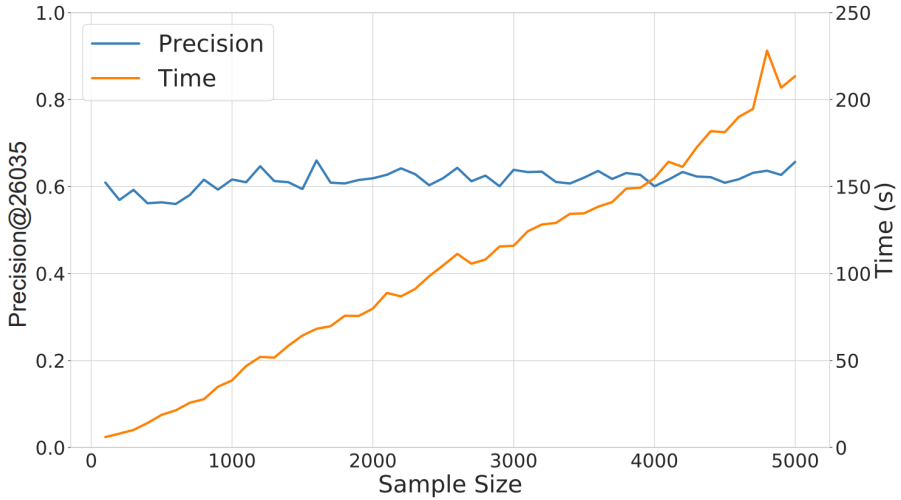
50

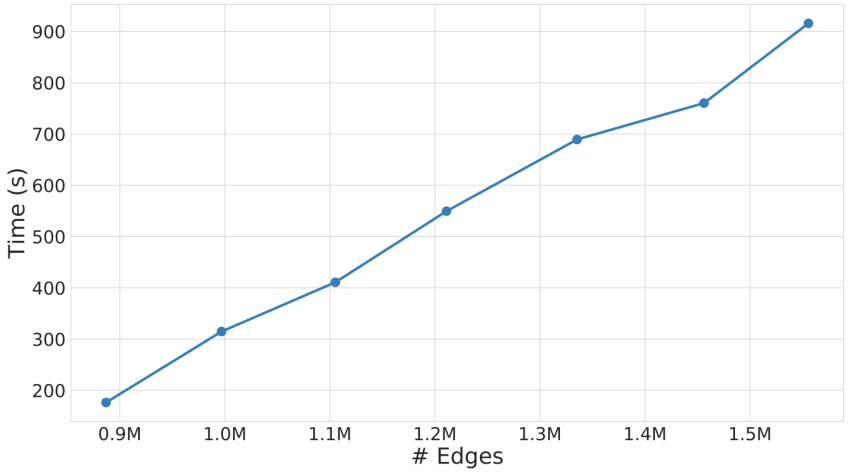
- Street Number or Identifier

Manitoba Hydro

- Data entry error

Experiments (Scalability)





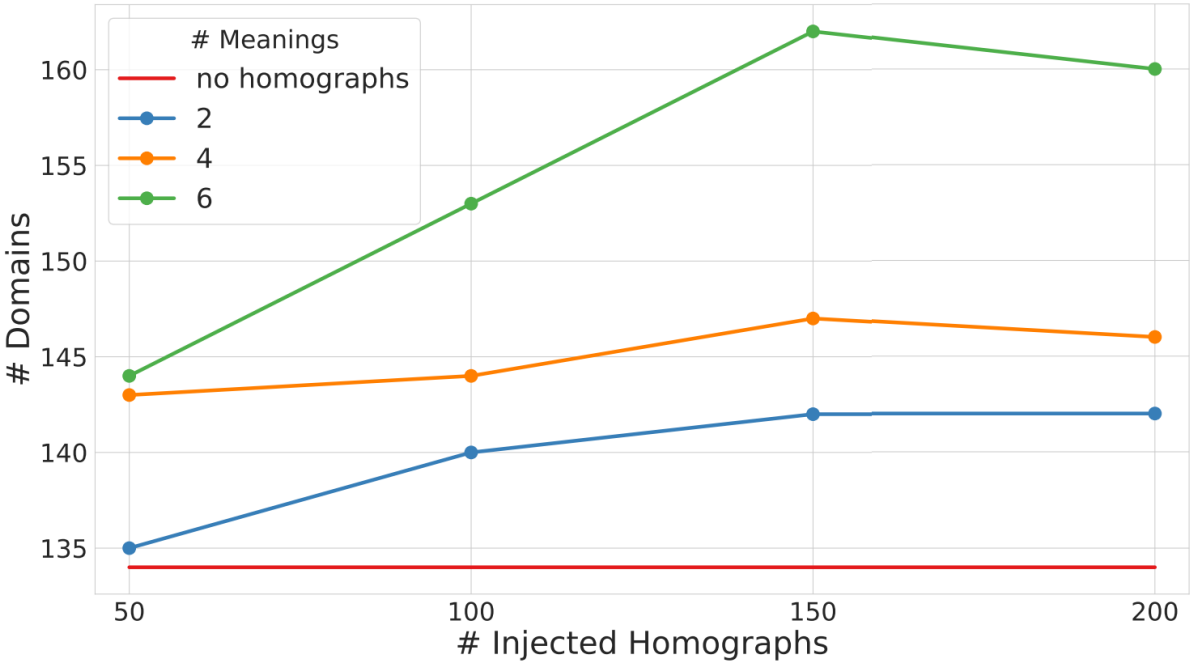
- Approximate BC [Geisberger+ 2008]
- Precision stabilizes even at 1% sampling of the nodes in the TUS benchmark
- Runtime grows linearly with the number of nodes sampled
  - $O(sm)$ time complexity of approximation

- Sample a various size subgraphs from the NYC Education dataset
- 27 minutes to approximate (1% sampling) BC scores for NYC Education
  - 1.5M nodes and 2.3M edges

R. Geisberger, P. Sanders, and D. Schultes. Better approximation of betweenness centrality. In ALENEX, pages 90–100. SIAM, 2008.

Experiments (Impact on D^4)

Homograph Injection: Select *n* data values from *n* non-unionable columns and replace them with a new unique value. This value is now a homograph with *n* meanings



- Knowledge of homographs can improve the performance of existing methods such as D^4
- 68 domains based on ground truth in the TUS-I dataset
- When homographs are removed D^4 comes closer to the ground truth

Conclusion and Future Work

DomainNet: unsupervised technique to identify if a data value is a homograph or not

- Examines data value co-occurrence
- Uses network-centrality measures on a bipartite graph representation of the data lake

Open Questions & Future Work

- Find number of meanings of a homograph
- Identify the meaning of an instance of an instance of a homograph
- Taxonomize homographs in data lakes
  - Null Values vs. misplaced values vs. *true* homographs
- Semantics benchmark for data lakes

Northeastern University
Khoury College of Computer Sciences

DATALAB @Northeastern

DomainNet: Homograph Detection for Data Lake Disambiguation

Aristotelis Leventidis, Laura Di Rocco, Wolfgang Gatterbauer, Renée J. Miller, Mirek Riedewald

Project Page: https://northeastern-datalab.github.io/table-as-query/
Code: https://github.com/northeastern-datalab/domain_net
Data Lab: https://db.khoury.northeastern.edu/

Northeastern University
Khoury College of Computer Sciences

DATALAB @Northeastern

DomainNet: Homograph Detection for Data Lake Disambiguation

Aristotelis Leventidis, Laura Di Rocco, Wolfgang Gatterbauer, Renée J. Miller, Mirek Riedewald

Project Page: https://northeastern-datalab.github.io/table-as-query/
Code: https://github.com/northeastern-datalab/domain_net
Data Lab: https://db.khoury.northeastern.edu/

An Example Data Lake

Col1	Col2	Col3
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

name_1	name_2	name_3
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

An Example Data Lake

Col1	Col2	Col3
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

name_1	name_2	name_3
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

An Example Data Lake

Col1	Col2	Col3
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

name_1	name_2	name_3
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

An Example Data Lake

Col1	Col2	Col3
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

name_1	name_2	name_3
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

An Example Data Lake

Col1	Col2	Col3
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

name_1	name_2	name_3
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

	0	1	2
Air China	CA	129	12:50
Malév Hungarian Airlines	MA	355	16:45
Air Madrid	NM	280	7:35

An Example Data Lake

Homograph: A data value in the data lake with more than one meaning

Col1	Col2	Col3
Pasadena	2016	240
Middlesex	2002	544
Luna	2004	256

name_1	name_2	name_3
Pasadena	CA	137122
Middlesex	MA	1611699
Luna	NM	23709

0	1	2
CA	129	12:50
MA	355	16:45
NM	280	7:35

DomainNet

Data Lake Disambiguation: Given a data lake containing a collection of tables with possibly missing, incomplete and/or heterogenous table and attribute names. Determine if a data value v that appears in more than one attribute or table has a single meaning or more than one meaning.

- Novel problem: Homograph detection in data lakes
- **DomainNet**: examines value co-occurrence and uses network-centrality measures on a bipartite graph representation of the data lake
- First benchmark (using real and synthetic data) for homograph detection in data lakes.

Disambiguation in Literature

Entity Resolution (ER)

- Do two tuples refer to the same real-world entity?
  - E.g., "X. Wang" one or many authors
- Resolution is between entities of the same type ☹
  - i.e., cannot resolve *Pasadena* the book versus the city

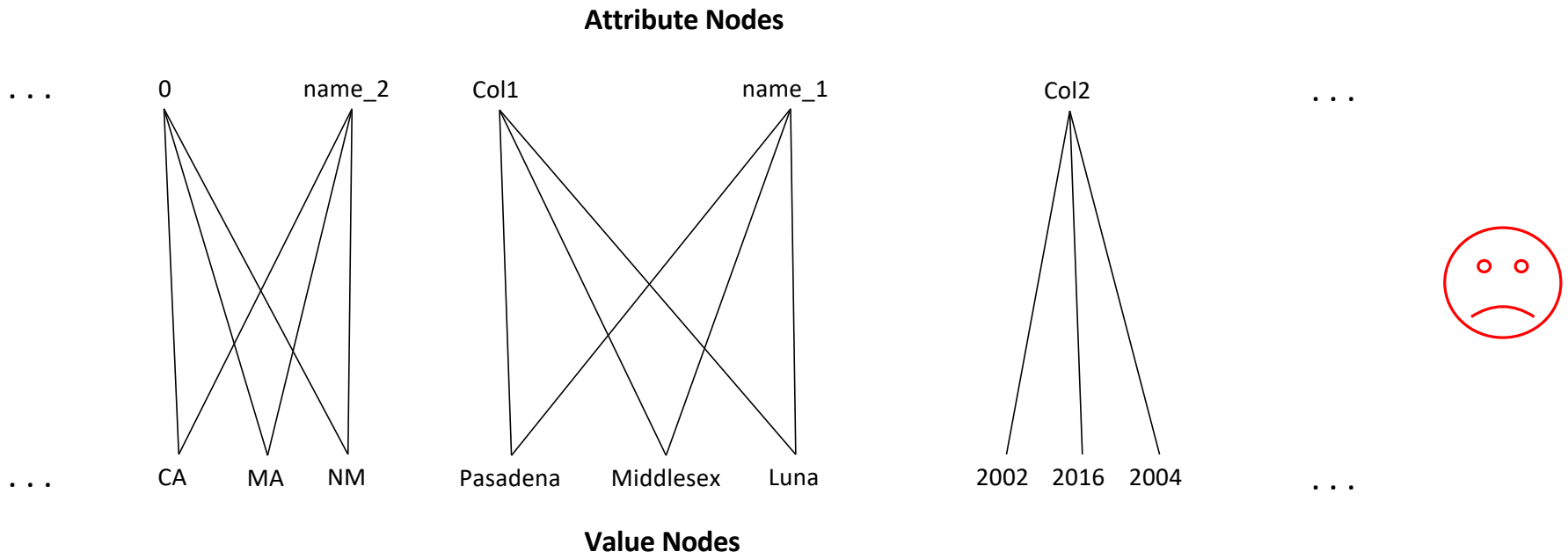
Semantic Type Detection

- Knowledge-based Techniques
  - ➢ Low coverage in data lakes
- Supervised Techniques
  - ➢ Sherlock, SATO
    - Limited number types (78)

- Unsupervised Techniques
  - ➢ D^4
    - Place data values to domains
    - We use it as a baseline

DomainNet (A graph Representation)

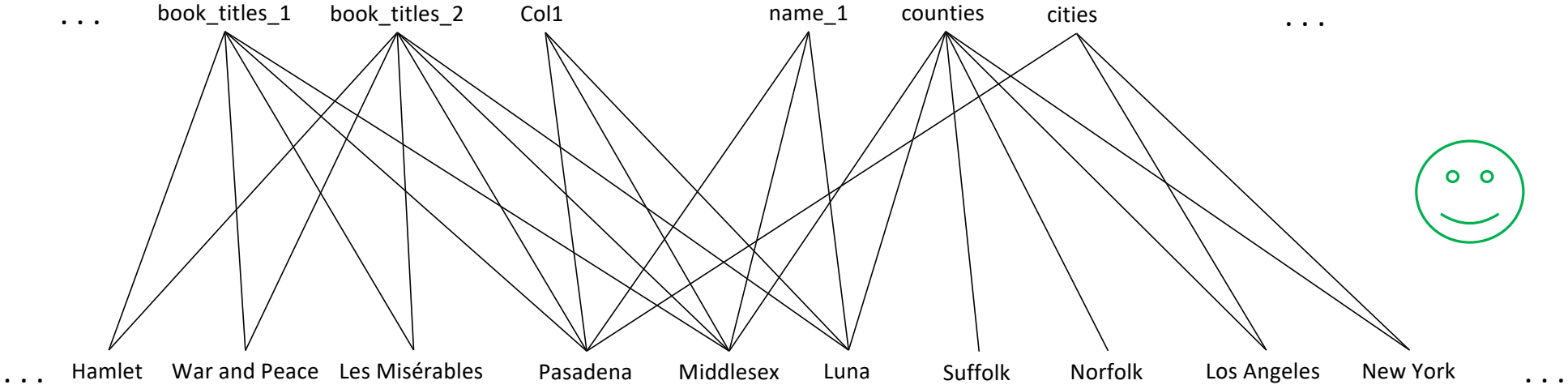
- Unsupervised
- A *homograph* likely co-occurs with a set of values that don't co-occur with each other
  - E.g, Book titles don't co-occur extensively with county names or cities.
- Bipartite graph (attribute nodes & value nodes)



DomainNet (A graph Representation)

- Unsupervised
- A *homograph* likely co-occurs with a set of values that don't co-occur with each other
  - E.g, Book titles don't co-occur extensively with county names or cities.
- Bipartite graph (attribute nodes & value nodes)

With enough data the co-occurrence pattern should emerge



Betweenness Centrality (BC)

- Construct a score for each data value that considers more than direct co-occurrences
- Betweenness centrality (BC) of a node measures how often a node lies on paths between all other nodes in the graph

$$BC(u) = \sum_{v \neq u,\, w \neq u} \frac{\sigma_{vw}(u)}{\sigma_{vw}}$$

Where σ_{vw} are the total number of shortest paths from v to w and $\sigma_{vw}(u)$ the number of shortest paths from v to w that pass through u

Hypothesis: A value node corresponding to a homograph will have a higher betweenness centrality than a value node with a single meaning

BC is expensive to compute exactly!

- $O(nm)$ time to compute where m is the number of edges and n the number of nodes

Use approximation techniques by sampling (Geisberger et al. sampling about 2-3% of the nodes)

- $O(sm)$ time to compute where s is the number of sample nodes chosen

Datasets

Dataset	#Tables	#Attributes	#Values	#Homographs
SB	13	39	17,633	55
TUS	1,327	9,859	190,399	26,035
TUS-I	1,253	5,020	163,860	N/A
NYC-EDU	201	3,496	1,469,547	N/A

Synthetic Benchmark (SB)

- Made using a data creator that specifies data sources (e.g. countries, car manufactures, animals, states etc.
- Each table has 1000 rows except table with countries and US states
- 55 homographs
  - Jaguar (car or animal), Lincoln (car or city), CA (country or state abbreviation)

Datasets

Dataset	#Tables	#Attributes	#Values	#Homographs
SB	13	39	17,633	55
TUS	1,327	9,859	190,399	26,035
TUS-I	1,253	5,020	163,860	N/A
NYC-EDU	201	3,496	1,469,547	N/A

Table Union Search (TUS) benchmark

- Repurpose a table unionability benchmark to provide a ground truth for homographs
- TUS made from real tables from open-data portals and provides for each column a set of other columns that it is unionable with.
- A data value is a homograph if it appears in at least two different columns that are not unionable.
- Significant portion of values are homographs (~13.7%)

Datasets

Dataset	#Tables	#Attributes	#Values	#Homographs
SB	13	39	17,633	55
TUS	1,327	9,859	190,399	26,035
TUS-I	1,253	5,020	163,860	N/A
NYC-EDU	201	3,496	1,469,547	N/A

TUS with Injected Homographs (TUS-I) benchmark

- Same as the TUS benchmark but with all the 163,860 homographs removed
- Used to artificially inject homographs in a controlled environment
  - Choose two different data values from columns that are not unionable and replace with a new unique value (e.g. "InjectedHomograph1")
- Vary number of meanings of the injected homographs
- Vary the cardinality of the replaced values
  - Cardinality of a data value v is the number of unique data values that it co-occurs with (#of 2-path-length neighbors)

Datasets

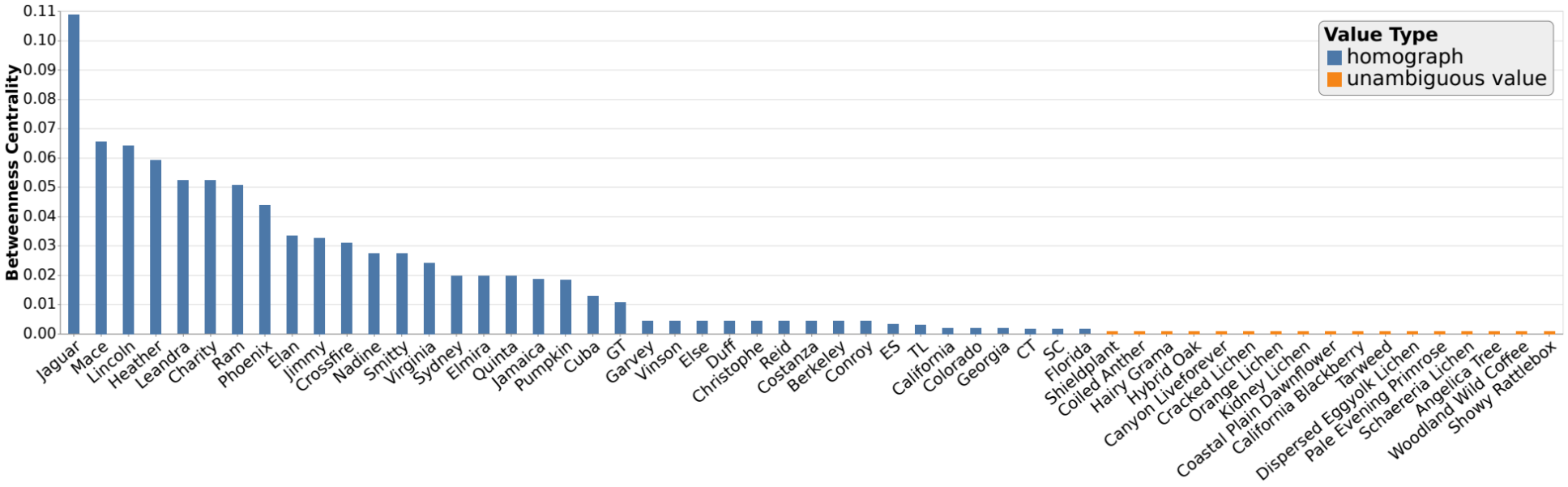
Dataset	#Tables	#Attributes	#Values	#Homographs
SB	13	39	17,633	55
TUS	1,327	9,859	190,399	26,035
TUS-I	1,253	5,020	163,860	N/A
NYC-EDU	201	3,496	1,469,547	N/A

New York City Education (NYC-EDU) benchmark

- Large repository of open data used to test the scalability of our method

Experiments (SB)

- Evaluation at top-55. There are 55 homographs in the SB based on ground truth
- 38/55≈69% are homographs vs. 38% achieved by D4
- Where are the remaining 17 homographs?
  - Remaining homographs correspond to country state name abbreviations (e.g. AL stands for Albania or Alabama)
  - Remaining homographs have low cardinality so proportionally fewer shortest paths pass through these homographs



Experiments (TUS-I)

- Approximate BC using 5000 nodes
- Select 50 pairs of data values from different columns and replace them with 50 injected homographs

Cardinality of replaced values	> 0	≥ 100	≥ 200	≥ 300	≥ 400	≥ 500
% of injected homographs @ top-50	85	93.5	93.5	95	94.5	97.5

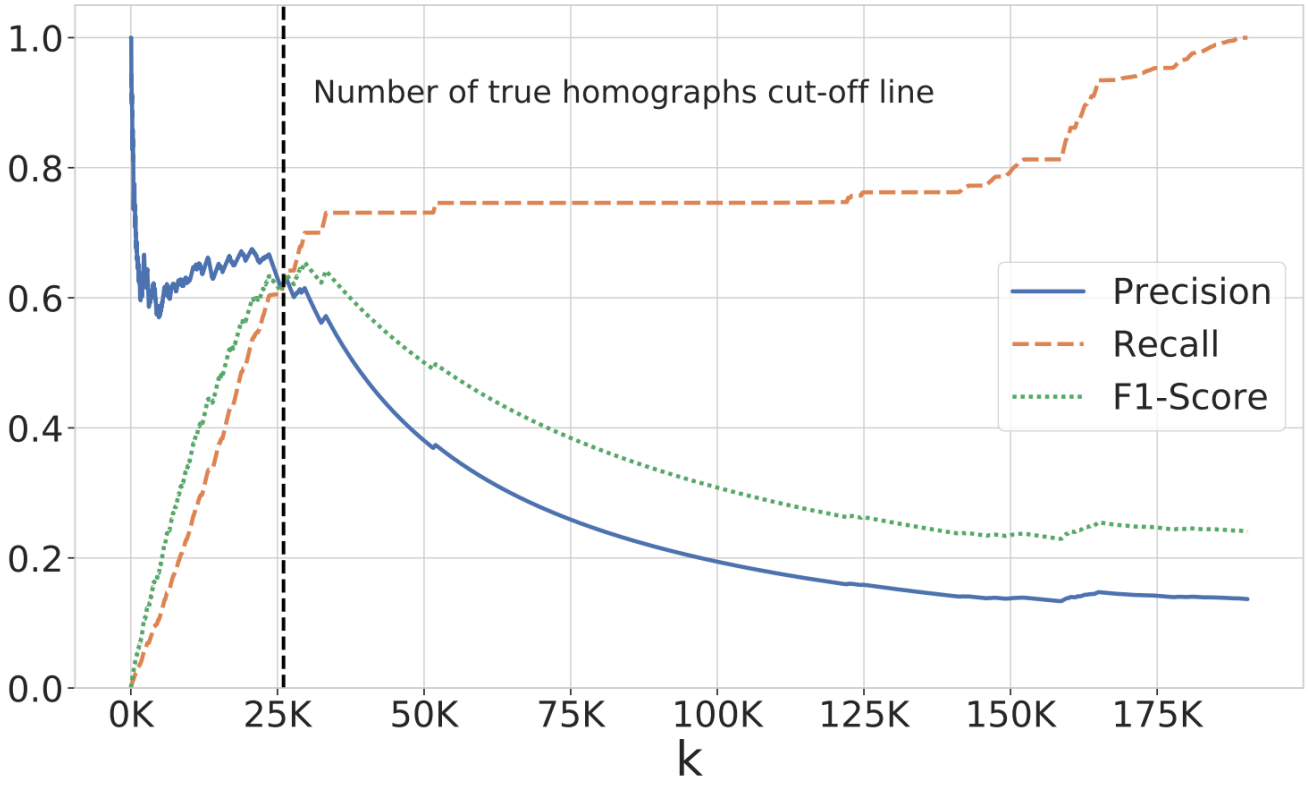
- Homographs with larger cardinality are more likely to be ranked higher (higher BC score)

- Select n data values from different columns and replace them with an injected homograph

# meanings of injected homographs	2	3	4	5	6	7	8
% of homographs @ top-50	97.5	97.5	98.5	100	100	100	100

- Homographs with more meanings are more likely to be ranked higher

Experiments (TUS)

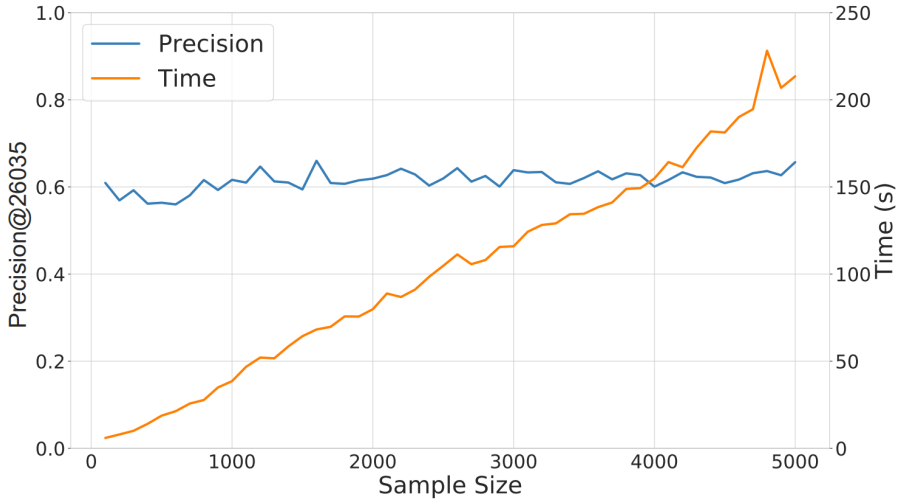


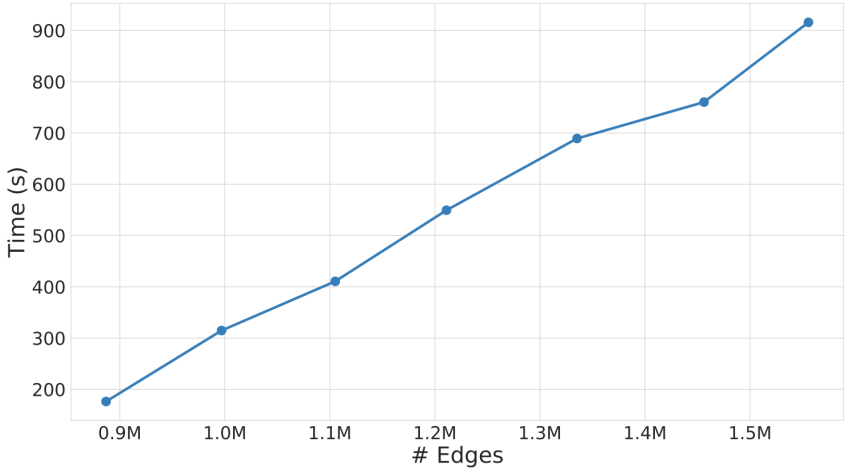
- Precision at k=200: 0.89
- Precision/Recall/F1-score at k=26035: 0.622

Some top values based on BC

- Music Faculty
- .
- 50
- Manitoba Hydro

Experiments (Scalability)

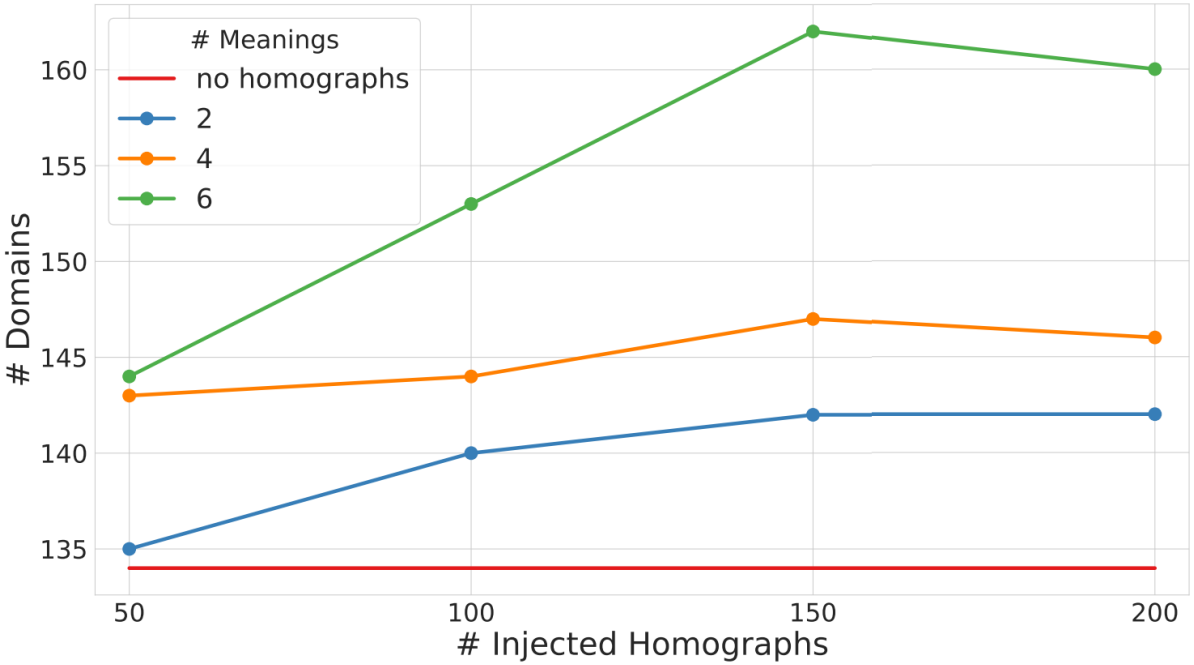




- Precision stabilizes even when sampling 1% of the nodes in the TUS benchmark
- Runtime grows linearly with the number of nodes sampled
  - $O(sm)$ time complexity of approximation

- Sample a various size subgraphs from the NYC Education dataset
- 27 minutes to approximate (1% sampling) BC scores for NYC Education
  - 1.5M nodes and 2.3M edges

Experiments (Impact on D^4)



- Knowledge of homographs can improve the performance of existing methods such as D^4.
- 68 domains based on ground truth in the TUS-I dataset
- When homographs are removed D^4 comes closer to the ground truth

Conclusion and Future Work

DomainNet: Unsupervised approach using betweenness centrality on a graph representation of the data to separate homographs from unambiguous values

- Number of meanings of homographs
- Identify the meaning of an instance of a homograph
- Taxonomy of homographs
  - Null values, misplaced values, true homographs
- Semantic benchmark for data lakes