High-quality Task Division for Large-scale Entity Alignment

Bing Liu[1], Wen Hua[1], Guido Zuccon[1], Genghong Zhao[2], Xia Zhang[2]

The University of Queensland[1], Neusoft[2]

bing.liu@uq.edu.au

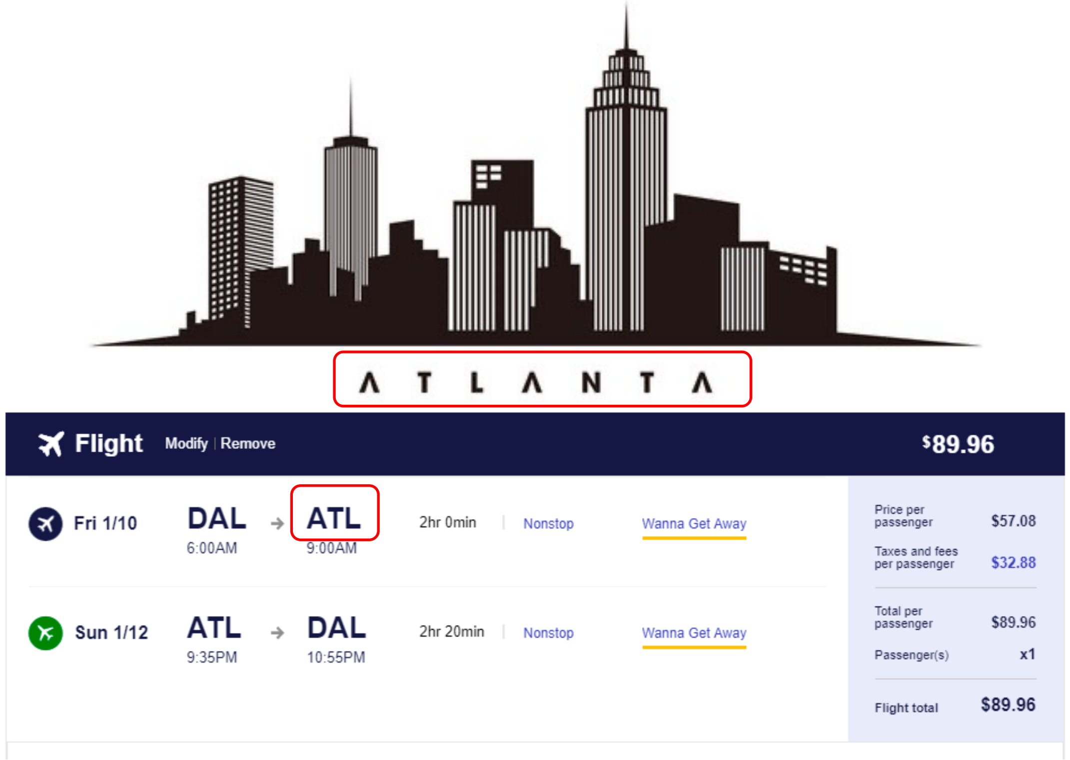
https://uqbingliu.github.io/

@BingLiu1011

Entity Alignment (EA)

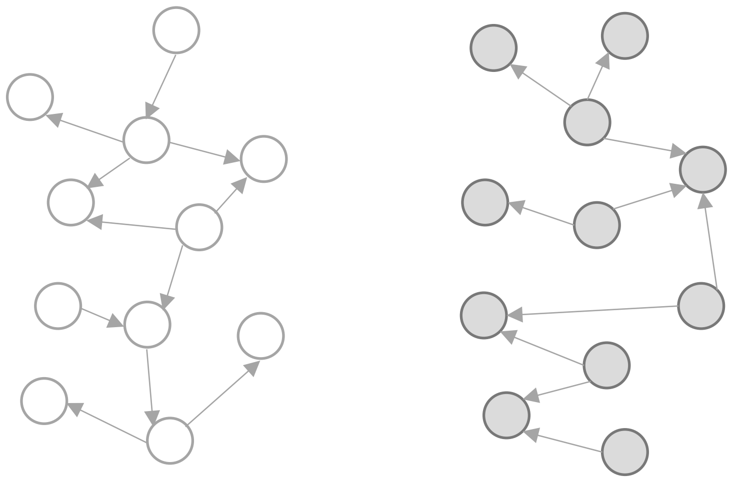
Entity Alignment aims to match **equivalent entities** in different **Knowledge Graphs** (KGs).

- One entity might be called differently in different scenarios.



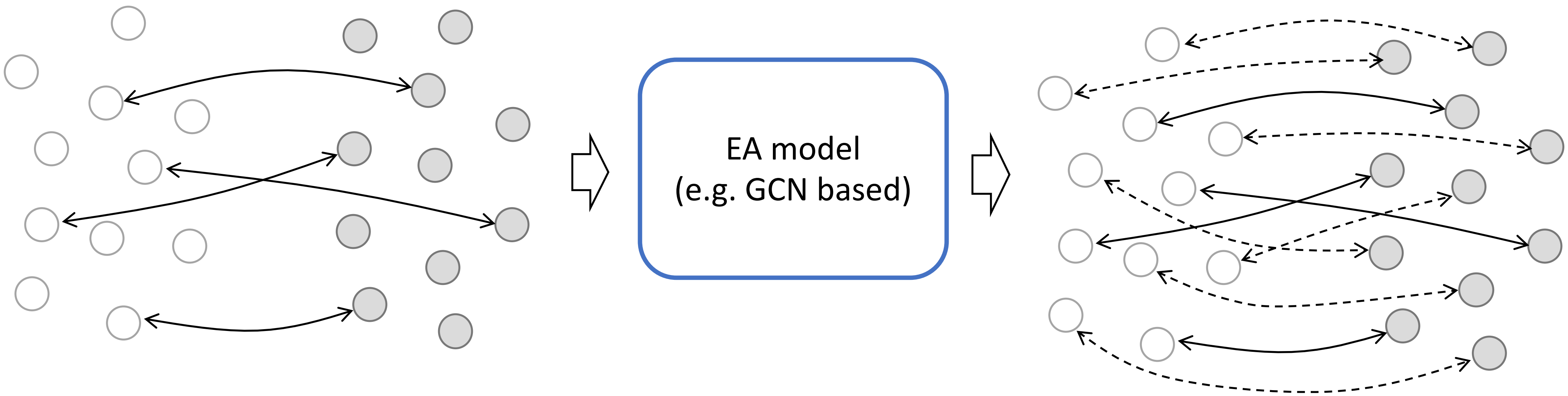
Entity Alignment (EA)

Entity Alignment aims to match **equivalent entities** in different **Knowledge Graphs** (KGs).



Neural Entity Alignment

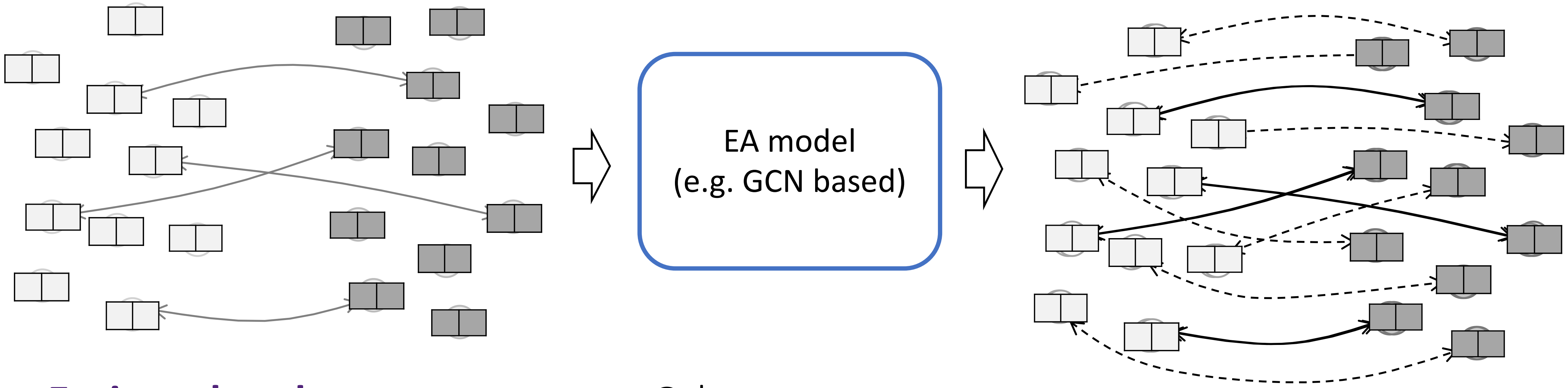
- Some **seed mappings** are provided as training data.
- Neural model encodes entities into **embeddings**.
- Predict **potential mappings**.



Scalability Issue

Neural EA models cannot be applied to large-scale KGs

- **Out-Of-Memory** (GPU).
- Time **efficiency**.



- **Entity-related parameters**
(initial entity representations)

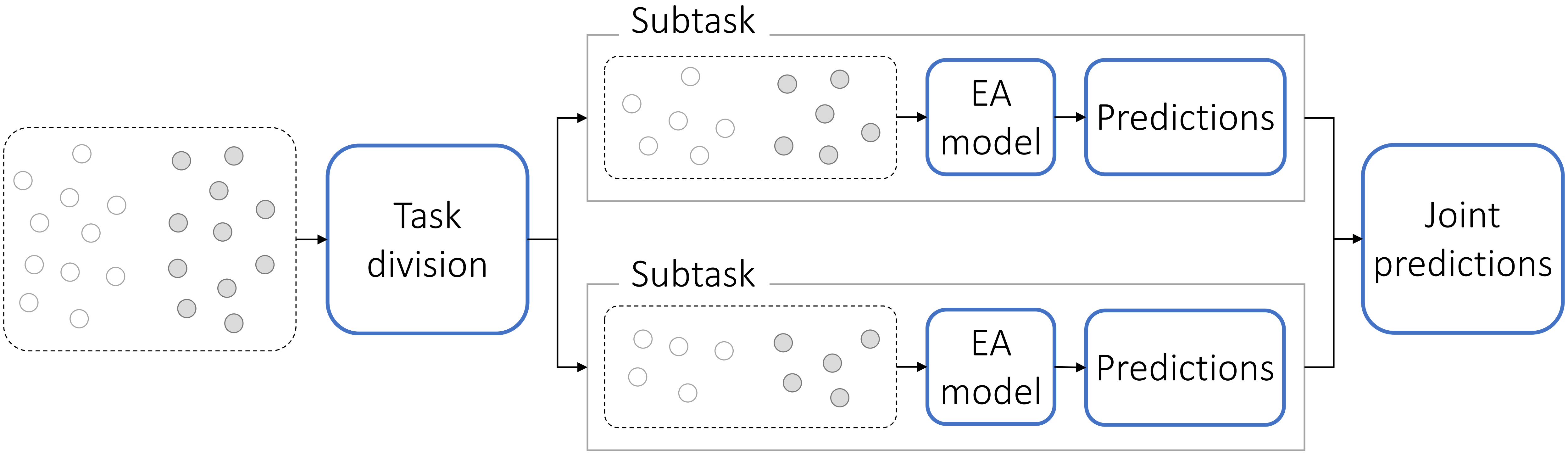
- Other parameters
- Neural operations

- **Entity representations**

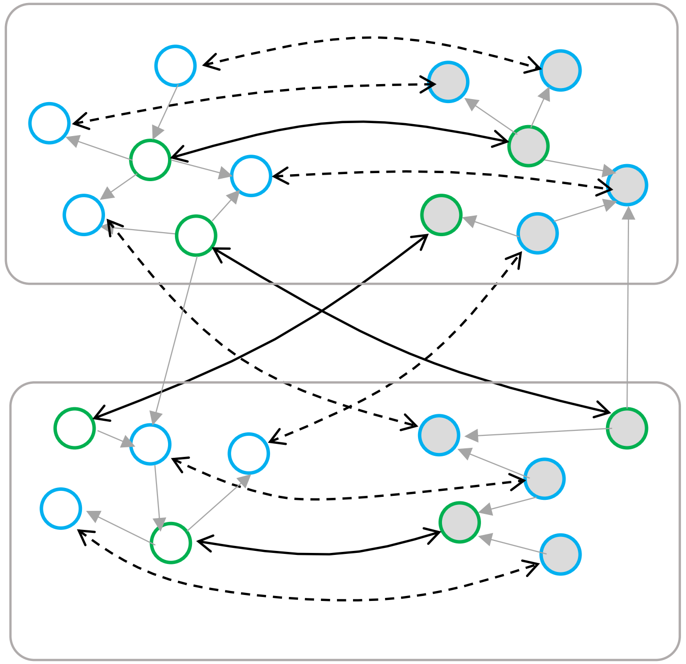
Task Division for Entity Alignment

Divide a large-scale EA task into multiple **small subtasks**

- Each subtask only has two **small subgraphs** to align.



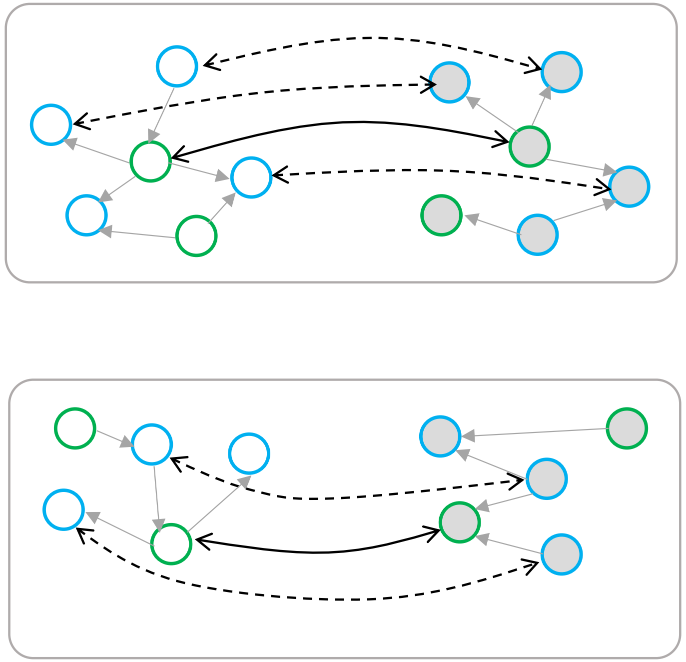
Challenges



Division

Lost info:

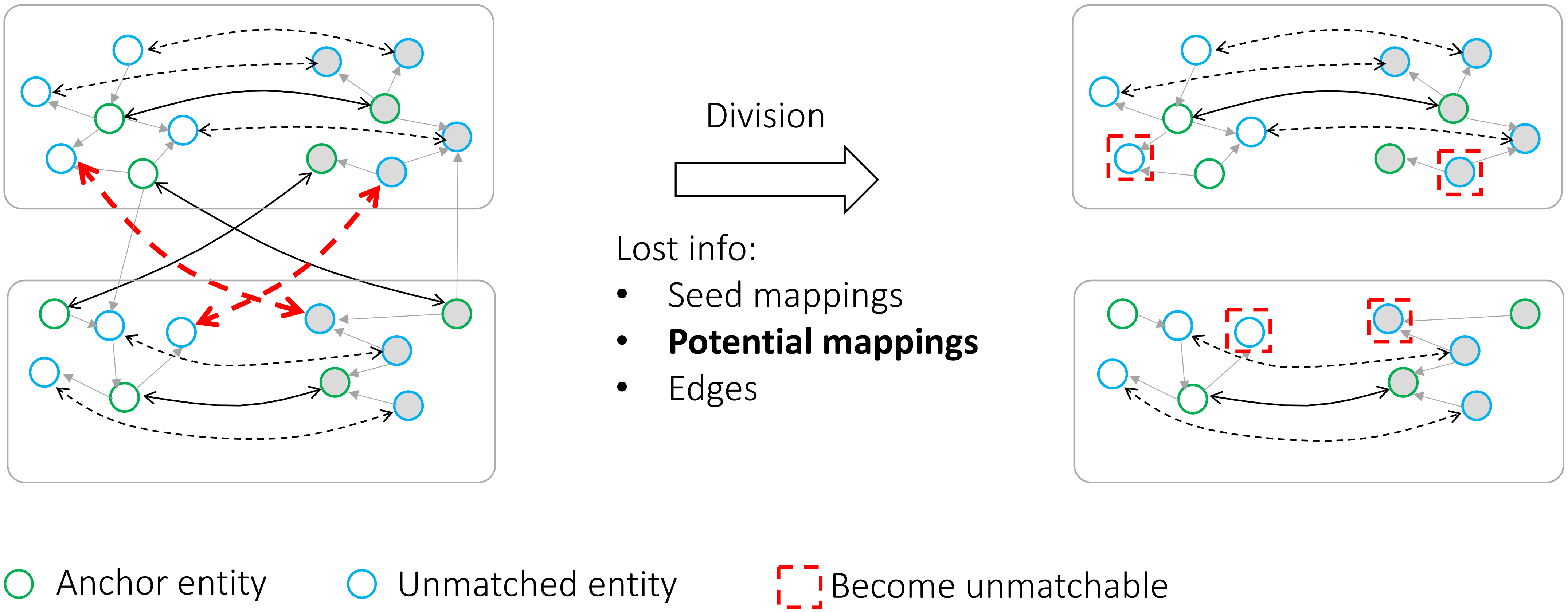
- **Seed mappings**
- **Potential mappings**
- **Edges**



Anchor entity   Unmatched entity

Challenges: Coverage of Potential Mappings

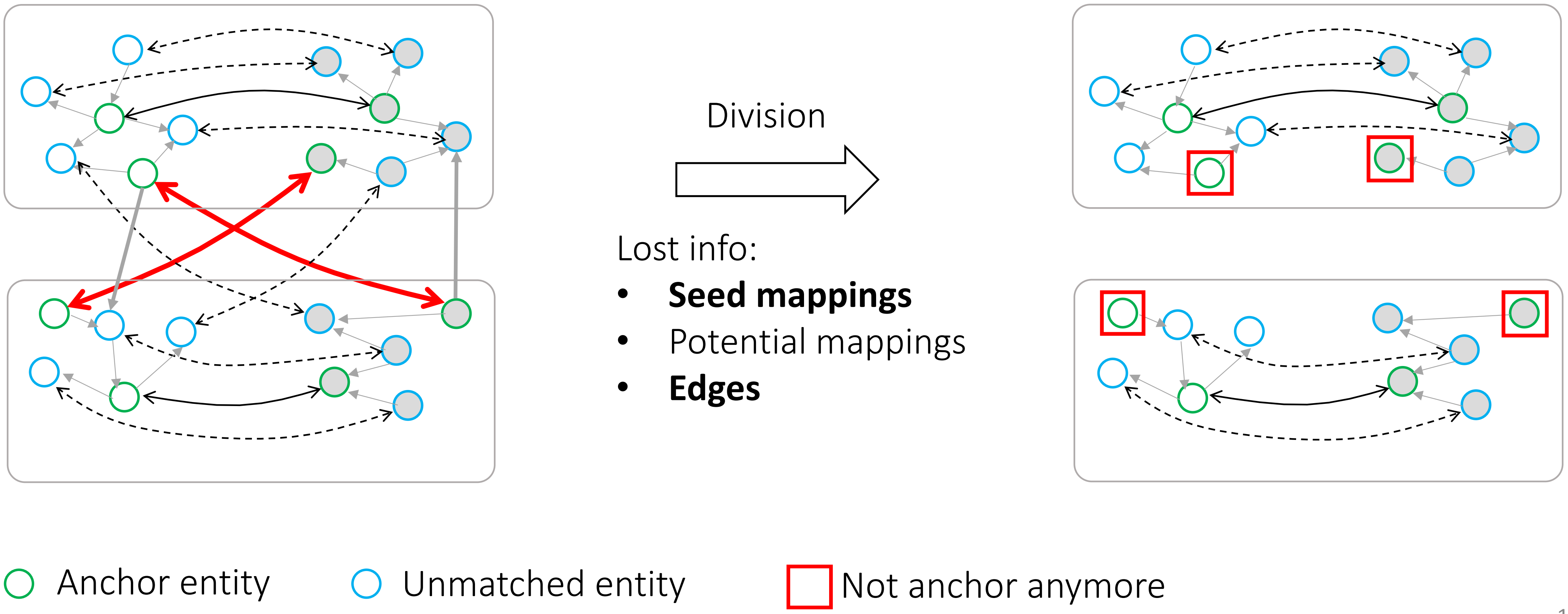
How to achieve high **coverage** of potential mappings?



Challenges: Informativeness of Context Graph

How to build informative **context graphs**?

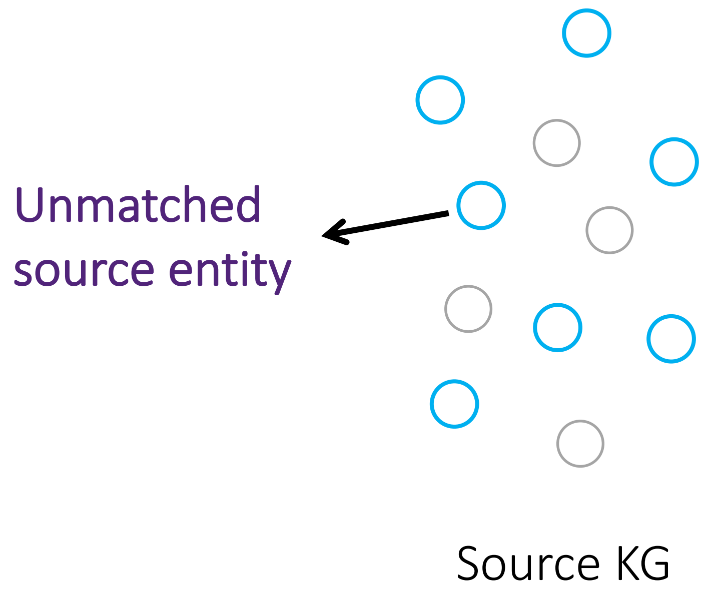
- The two graphs that contain the unmatched entities and are fed into the EA model. They provide **evidence** for entity matching.



The DivEA Framework

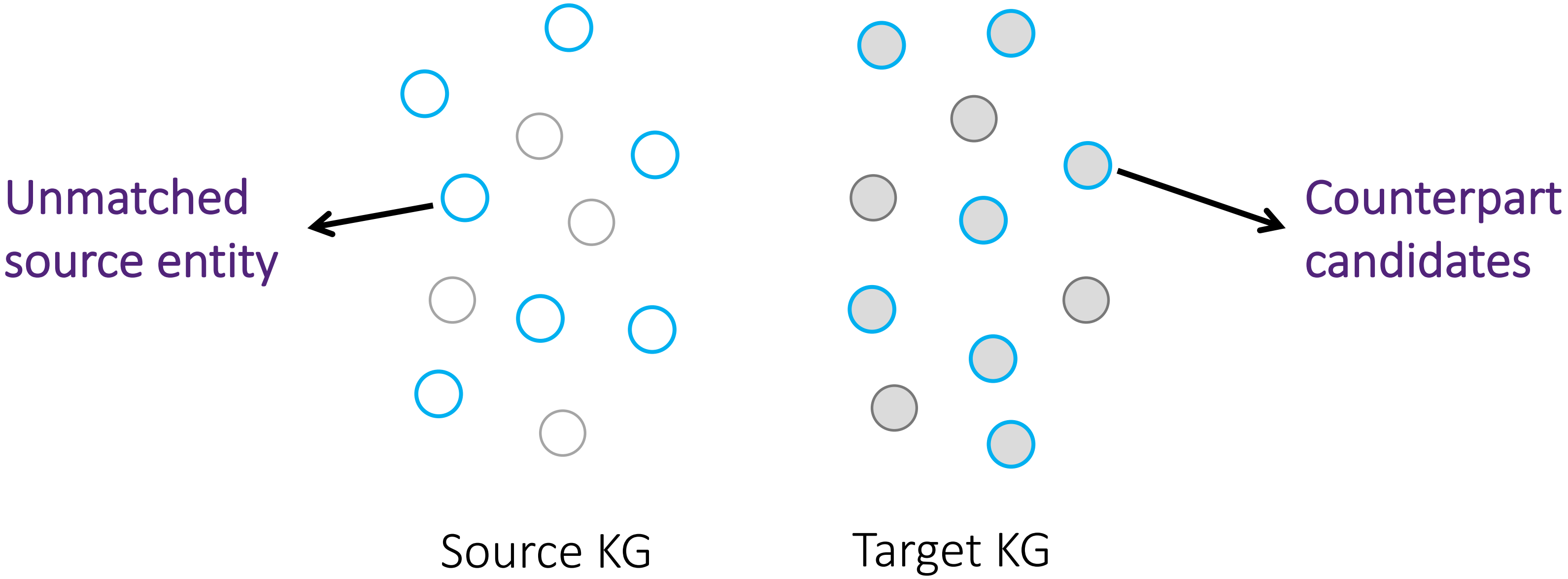
Elements of Entity Alignment

- **Unmatched source entities**



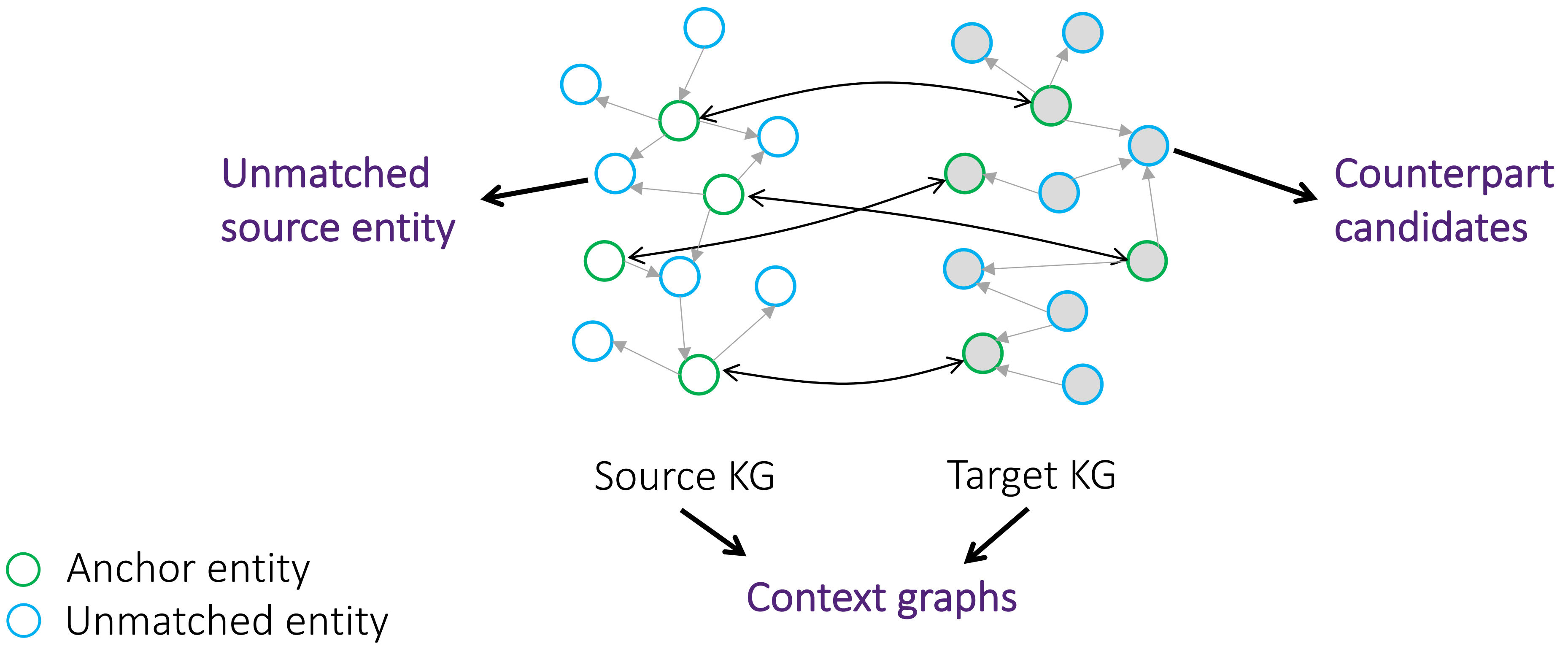
Elements of Entity Alignment

- **Unmatched source entities**
- **Counterpart candidates,** i.e. unmatched target entities



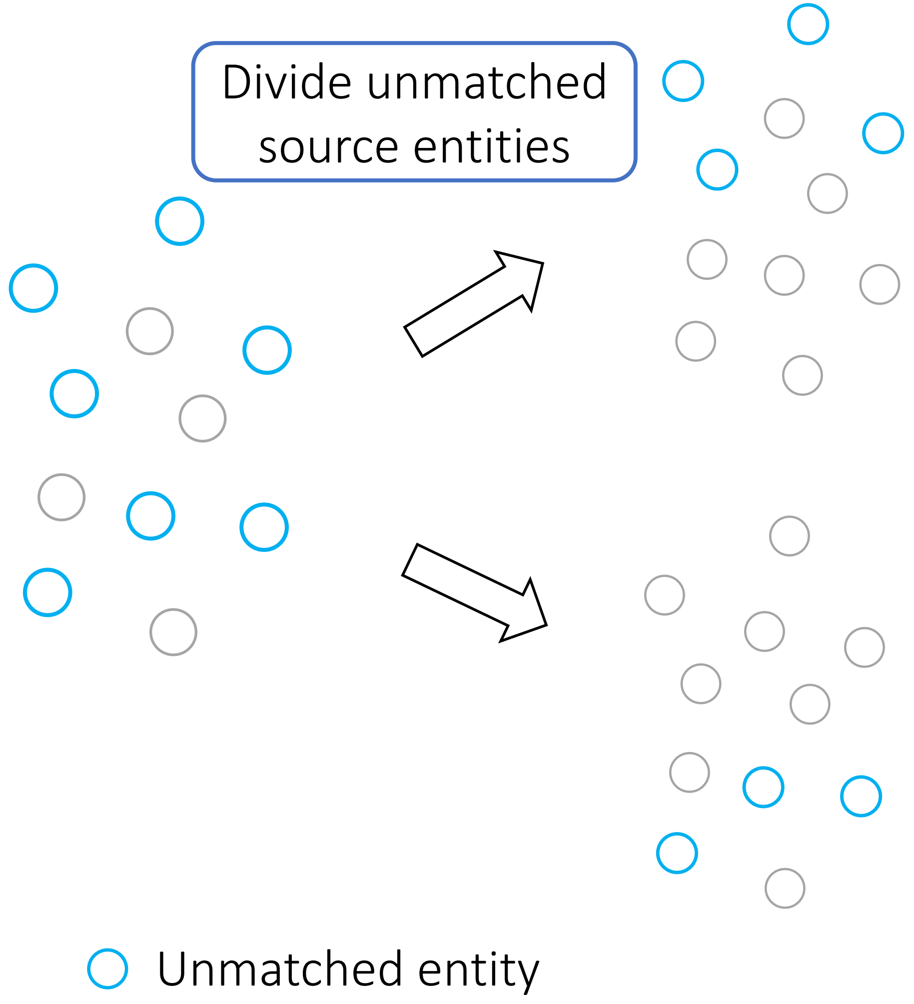
Elements of Entity Alignment

- **Unmatched source entities**
- **Counterpart candidates**, i.e. unmatched target entities
- **Context graphs**



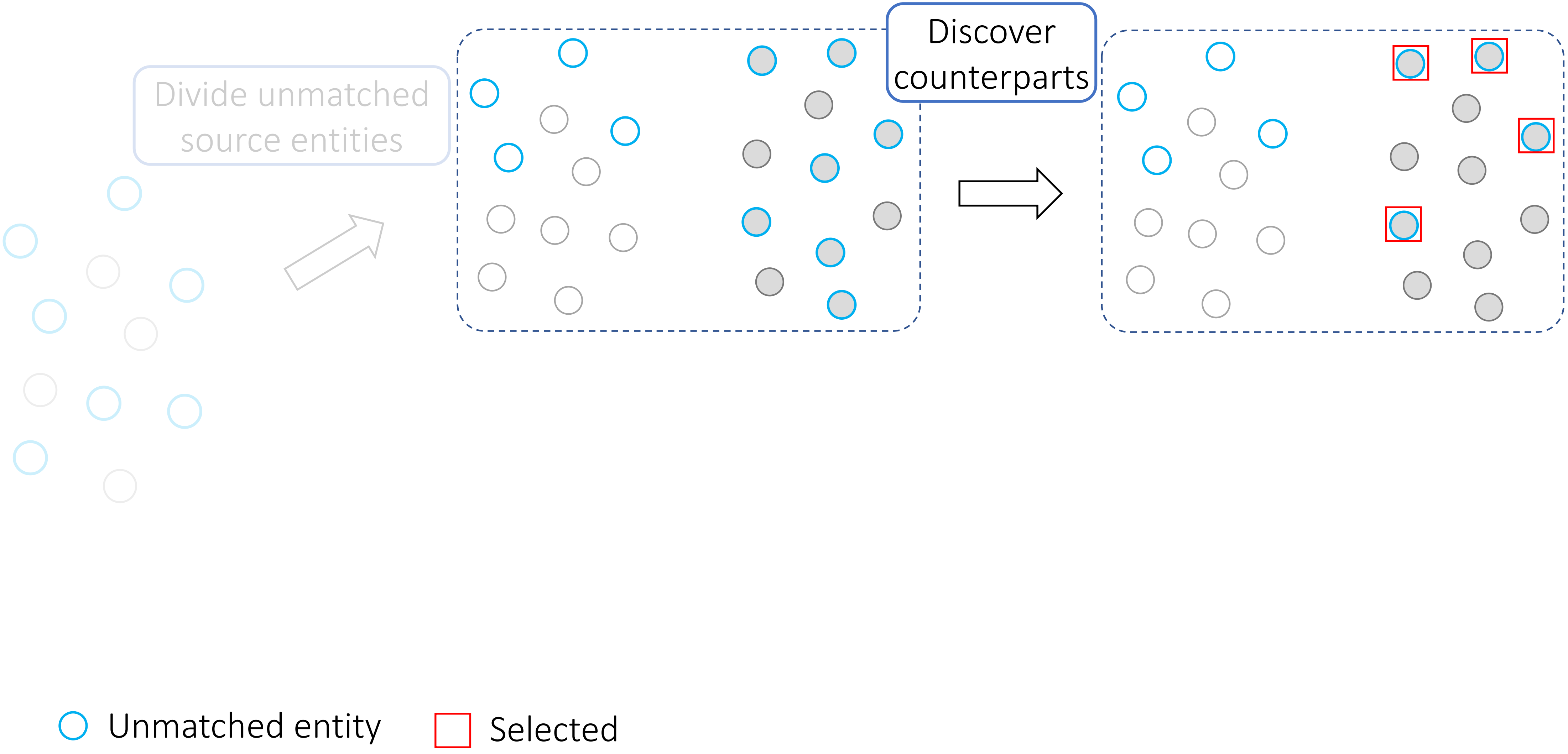
Overview: Divide Unmatched Source Entities

- Divide unmatched source entities



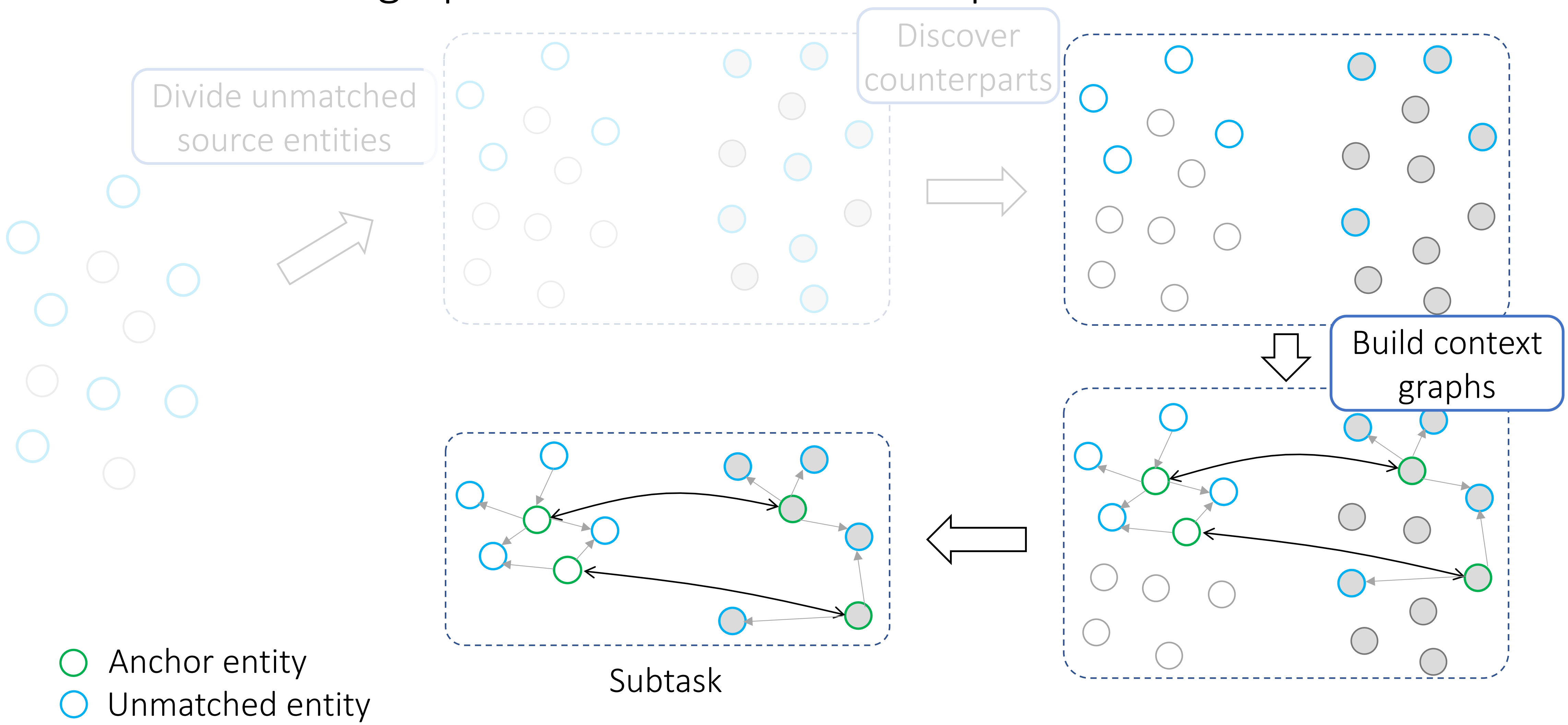
Overview: Counterpart Discovery

- Counterpart discovery: select a limited number of target entities



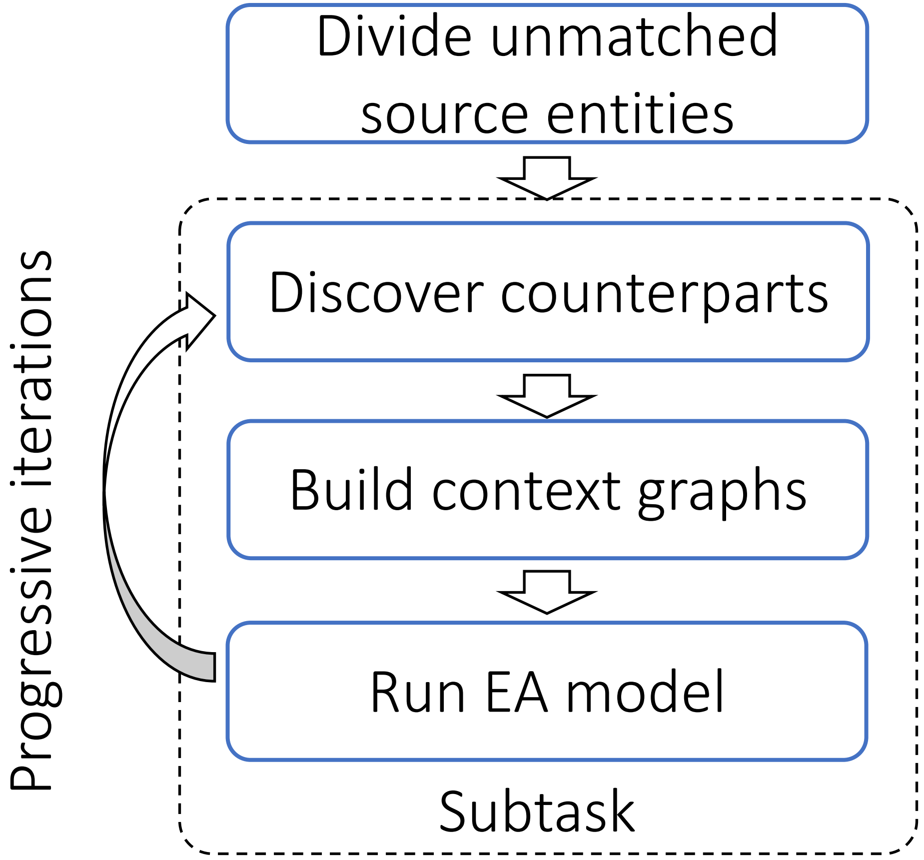
Overview: Build Context Graphs

- Build context graphs: add more entities to provide evidence.



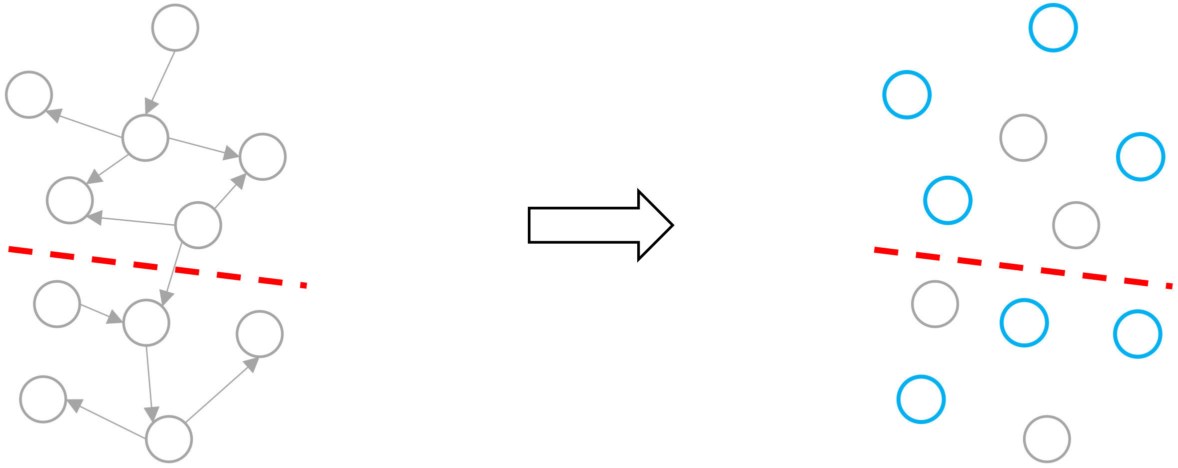
Overview: Overall Process

- Progressive process



Dividing Unmatched Source Entities

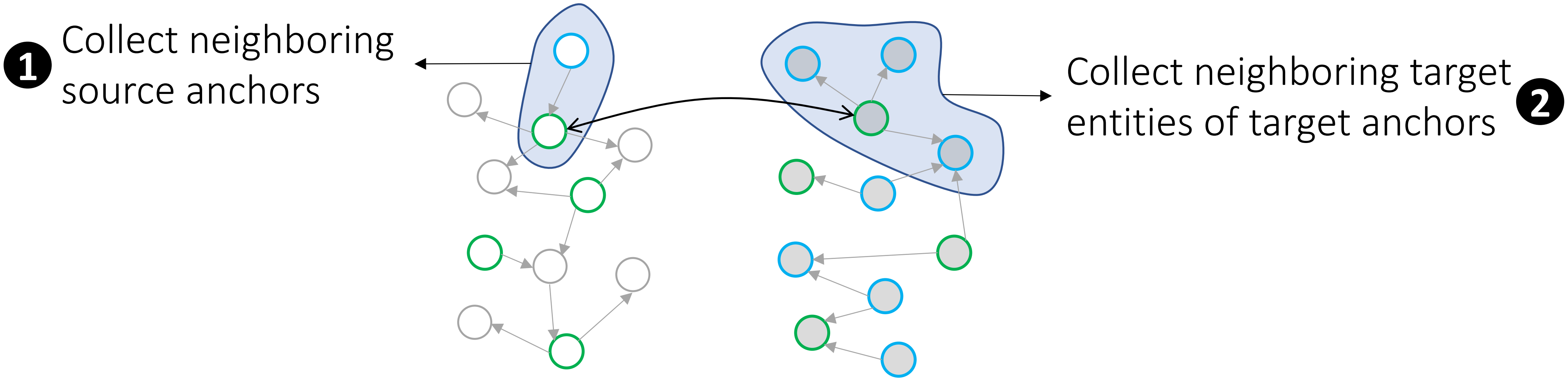
- Partition the source KG into **cohesive subgraphs** using *Metis*
  - The least cut-off of edges.
  - Balanced sizes.
- The unmatched source entities in each partition form one subset.



Counterpart Discovery: Principle of Locality

Given a certain source entity, how to identify its potential counterpart without using EA model?

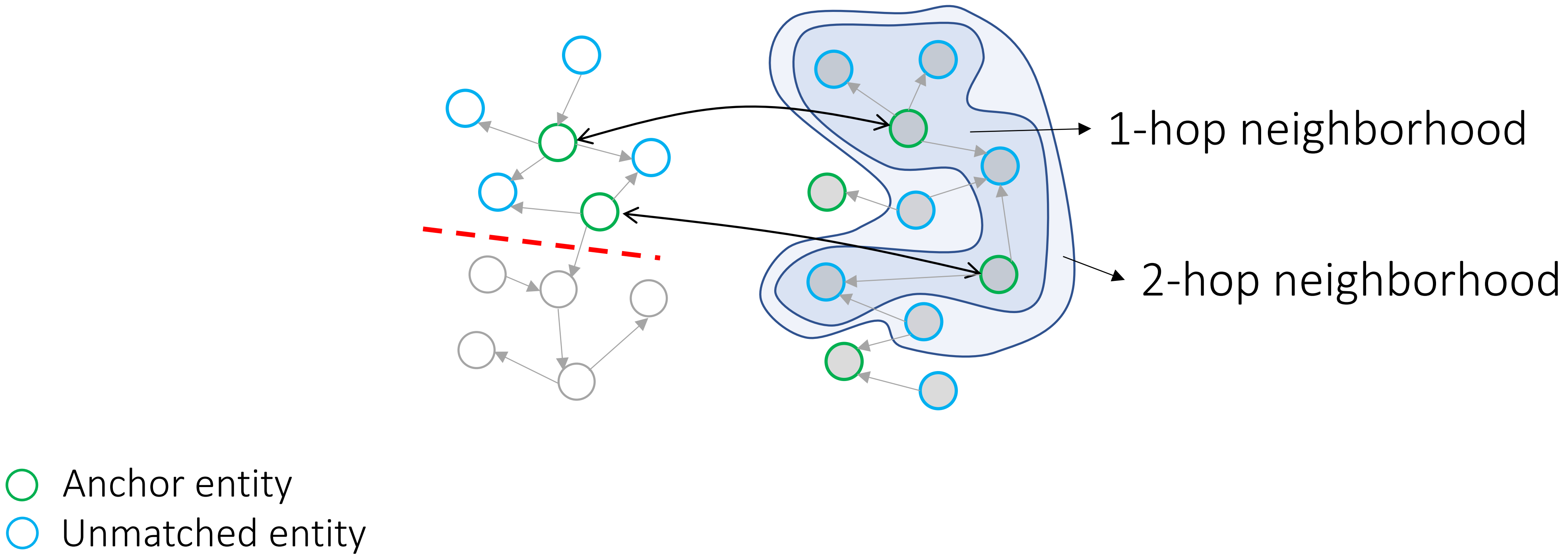
- Principle of **locality**: If two entities are equivalent, the other entities semantically related to them might also be equivalent.



Anchor entity
Unmatched entity

Counterpart Discovery: Principle of Locality

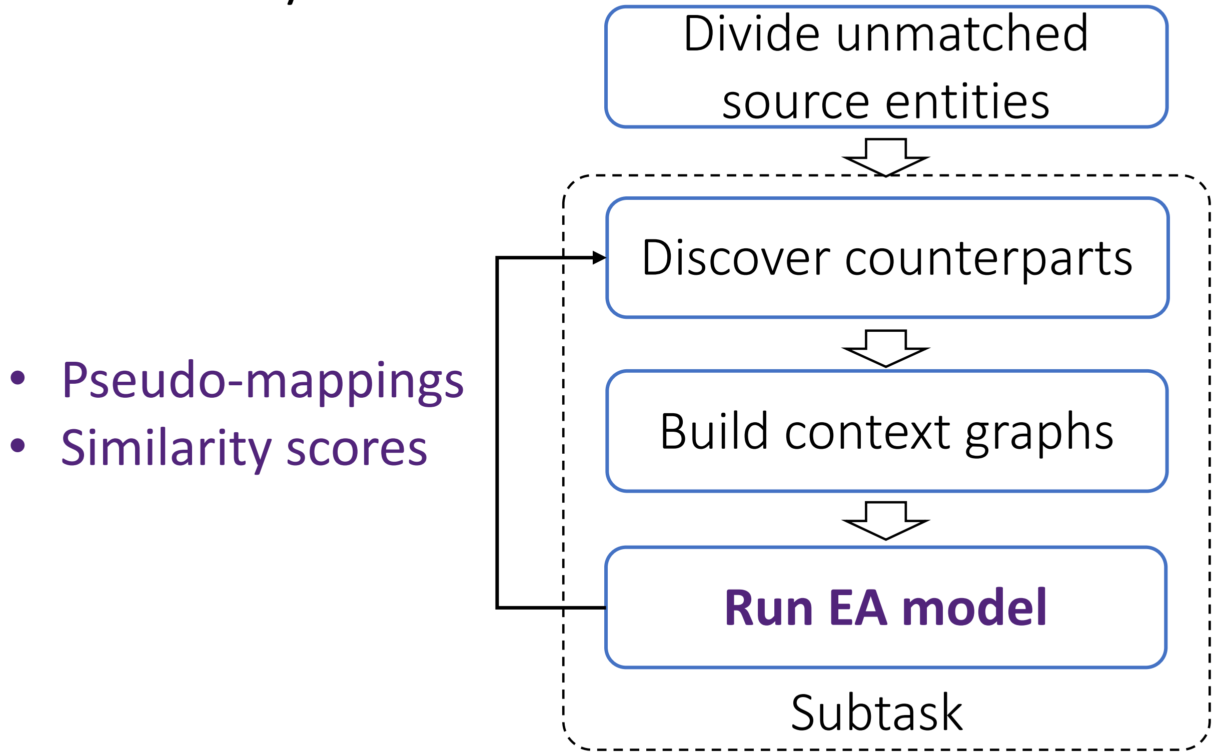
1. Collect anchors in the same graph partition.
2. Locality-based weight $W^{loc}(e^t)$ according to the distance between target entity e^t and target anchors.



Counterpart Discovery: Enhancement with EA Model

If you have an EA model, how to use it for counterpart discovery?

- Enrich the seed mappings.
- Similarity-based signal $W^{sim}(e^t)$ indicating the likelihood that e^t is the counterpart of any source entity.



Counterpart Discovery

Choose target entities with the highest overall weights $W(e^t)$

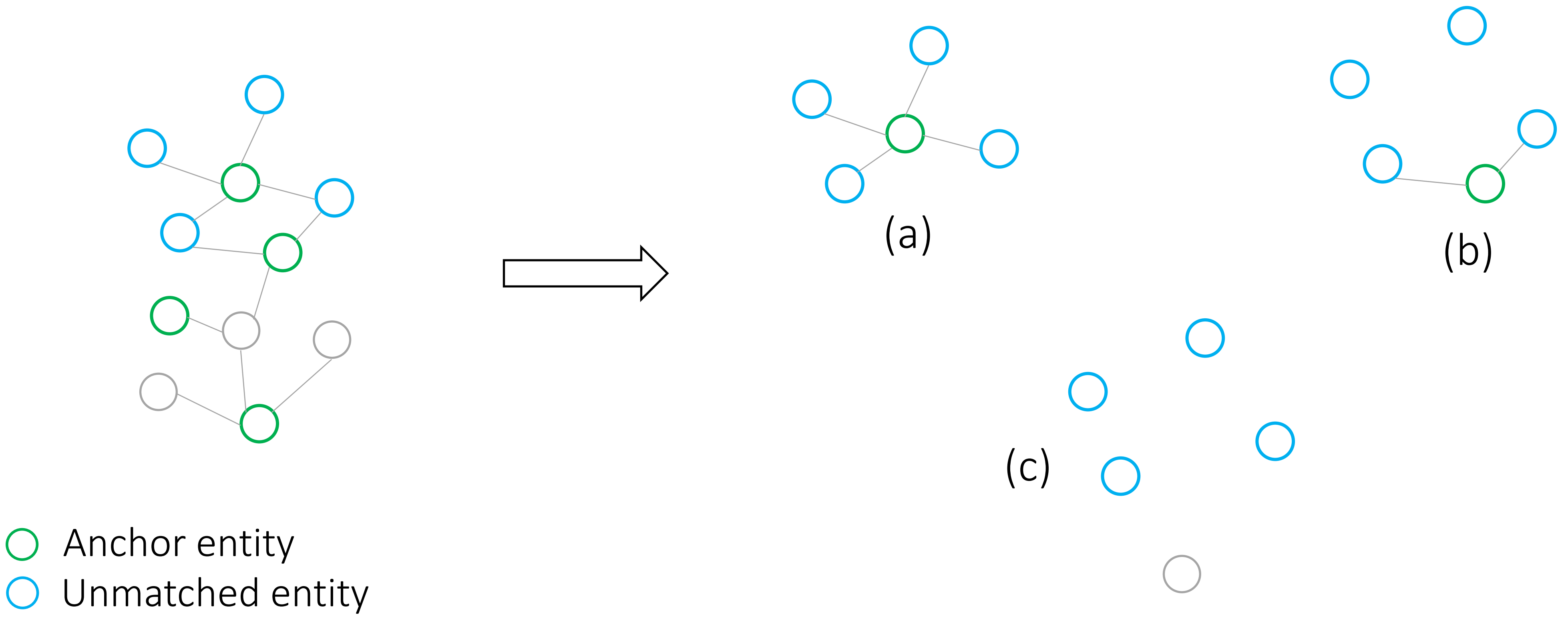
- β is a hyper-parameter.

$$W(e^t) = W^{loc}(e^t) + \beta\, W^{sim}(e^t)$$

Building Context Graphs: The Context Matters

The context graph matters a lot for EA model.

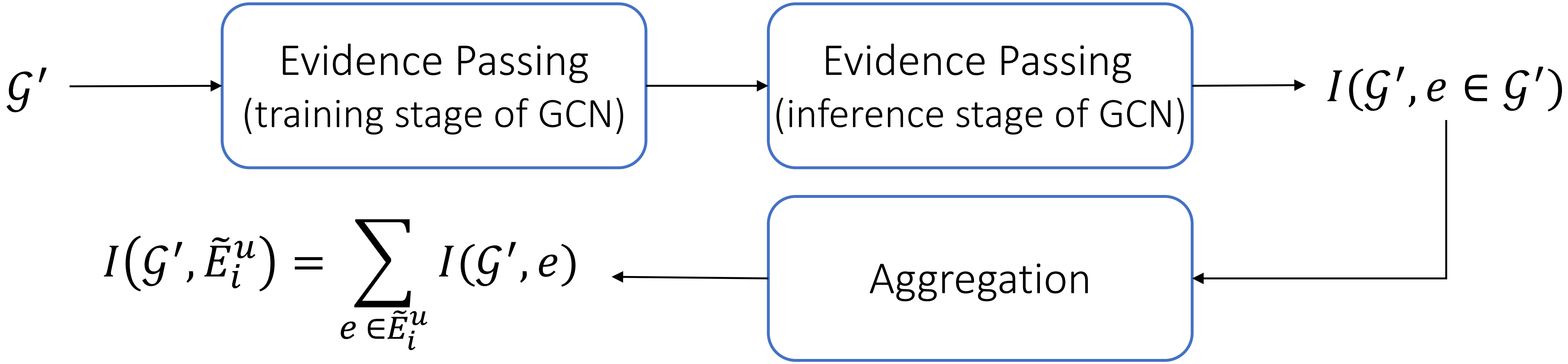
- Example: build a context graph of size 5 for the unmatched entities.
- The unmatched entities can get different evidence.



Building Context Graphs: Quantifying Informativeness

How to quantify the informativeness of a context graph?

- **Evidence Passing** mechanism: to simulate how evidence spread around a graph in a GCN-based EA model.
    - The evidence is **scalar** instead of high-dimension vector.
    - The evidence origins from the anchors.
    - The evidence spreads in the **training and inference stages** of GCN.



Building Context Graphs

- With the quantification method, we can search the most informative context graph within a single KG.

- For a subtask, we build the source context graph first, and then the target one.

Experiments

Experimental Setup

The University of Queensland Australia

Comparison with 2 baselines: CPS, SBP.

Task division for 2 EA models: GCN-Align, RREA

Evaluated on 6 datasets: DBP15K: FR-EN, JA-EN, ZH-EN; DWY100K: DBP-WD, DBP-YG; FB-DBP (2M)

Metrics: Hit@1 (H@1), Hit@5 (H@5), Mean Reciprocal Rank (MRR)

Results: Comparison with Baselines (CPS variants)

- Overall performance

Method	EA model	FR-EN (15K)		
		H@1	H@5	MRR
CPS (sup)	GCN-Align	0.151	0.396	0.263
CPS (semi)		0.274	0.478	0.367
DivEA		**0.396**	**0.642**	**0.504**
CPS (sup)	RREA	0.419	0.631	0.514
CPS (semi)		0.516	0.682	0.590
DivEA		**0.645**	**0.795**	**0.711**

Results: Comparison with Baselines (CPS variants)

- Overall performance

Method	EA model	FR-EN (15K)			FB-DBP (2M)		
		H@1	H@5	MRR	H@1	H@5	MRR
CPS (sup)	GCN-Align	0.151	0.396	0.263	0.000	0.000	0.000
CPS (semi)	GCN-Align	0.274	0.478	0.367	0.000	0.000	0.000
DivEA	GCN-Align	**0.396**	**0.642**	**0.504**	**0.051**	**0.106**	**0.08**
CPS (sup)	RREA	0.419	0.631	0.514	0.043	0.080	0.062
CPS (semi)	RREA	0.516	0.682	0.590	0.056	0.089	0.073
DivEA	RREA	**0.645**	**0.795**	**0.711**	**0.163**	**0.24**	**0.202**

Results: Comparison with Baselines (SBP variants)

- Overall performance

Method	EA model	FR-EN (15K)			FB-DBP (2M)		
		H@1	H@5	MRR	H@1	H@5	MRR
SBP (sup)	GCN-Align	0.163	0.426	0.284	0.000	0.000	0.000
SBP (semi)		0.288	0.511	0.391	0.005	0.011	0.008
I-SBP		0.175	0.372	0.267	0.000	0.000	0.000
DivEA		**0.402**	**0.678**	**0.525**	**0.071**	**0.15**	**0.112**
SBP (sup)	RREA	0.475	0.721	0.583	0.070	0.139	0.106
SBP (semi)		0.575	0.762	0.659	0.095	0.159	0.128
I-SBP		0.508	0.730	0.608	0.120	0.233	0.172
DivEA		**0.655**	**0.841**	**0.736**	**0.199**	**0.298**	**0.248**

Results: Comparison with Baselines

- Coverage of potential mappings
  - Metric: recall of potential mappings in the subtasks

	15K			100K		2M
Method	FR-EN	JA-EN	ZH-EN	DBP-WD	DBP-YG	FB-DBP
CPS	0.817	0.718	0.826	0.542	0.486	0.237
DivEA	**0.881**	**0.892**	**0.880**	**0.830**	**0.893**	**0.507**

Results: Comparison with Baselines

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

- Coverage of potential mappings
  - Metric: recall of potential mappings in the subtasks

	15K			100K		2M
Method	FR-EN	JA-EN	ZH-EN	DBP-WD	DBP-YG	FB-DBP
CPS	0.817	0.718	0.826	0.542	0.486	0.237
DivEA	**0.881**	**0.892**	**0.880**	**0.830**	**0.893**	**0.507**

Results: Comparison with Baselines

The University of Queensland Australia

- Coverage of potential mappings
  - Metric: recall of potential mappings in the subtasks

	15K			100K		2M
Method	FR-EN	JA-EN	ZH-EN	DBP-WD	DBP-YG	FB-DBP
CPS	0.817	0.718	0.826	0.542	0.486	0.237
DivEA	**0.881**	**0.892**	**0.880**	**0.830**	**0.893**	**0.507**
SBP	0.930	0.942	0.943	0.819	0.824	0.426
I-SBP	0.960	0.957	0.960	0.947	0.982	0.502
DivEA	**0.978**	**0.979**	**0.970**	**0.954**	**0.994**	**0.684**

Larger subtask size

Results: Comparison with Baselines

The University of Queensland Australia

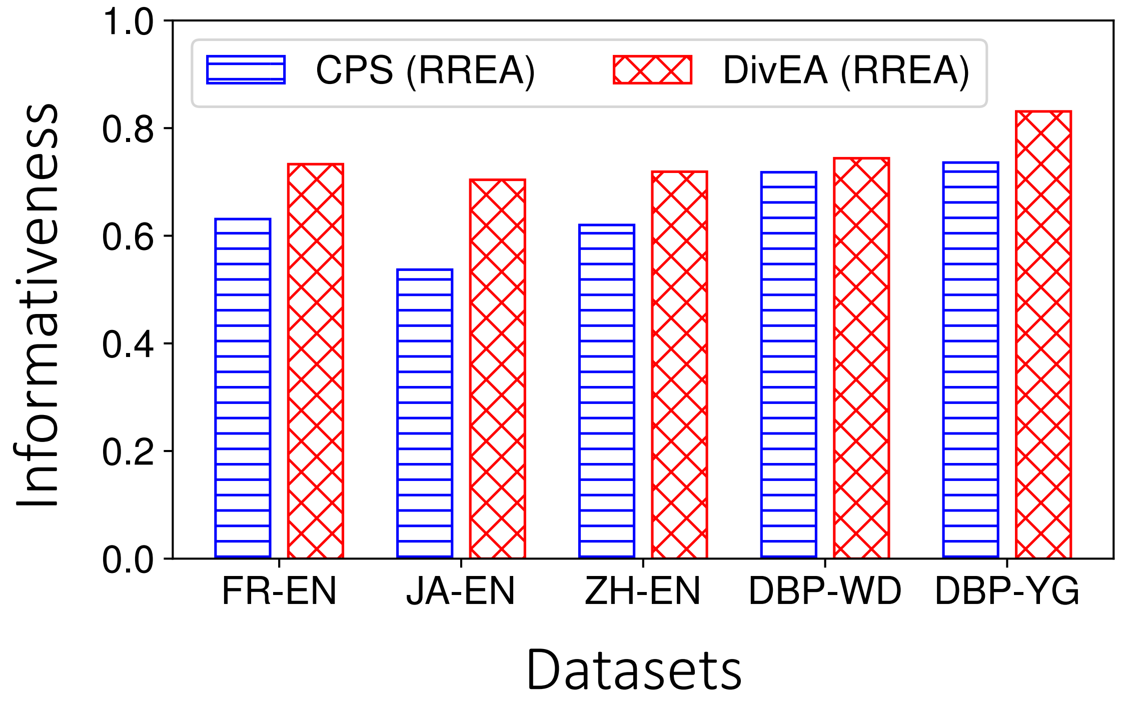
- Coverage of potential mappings
  - Metric: recall of potential mappings in the subtasks

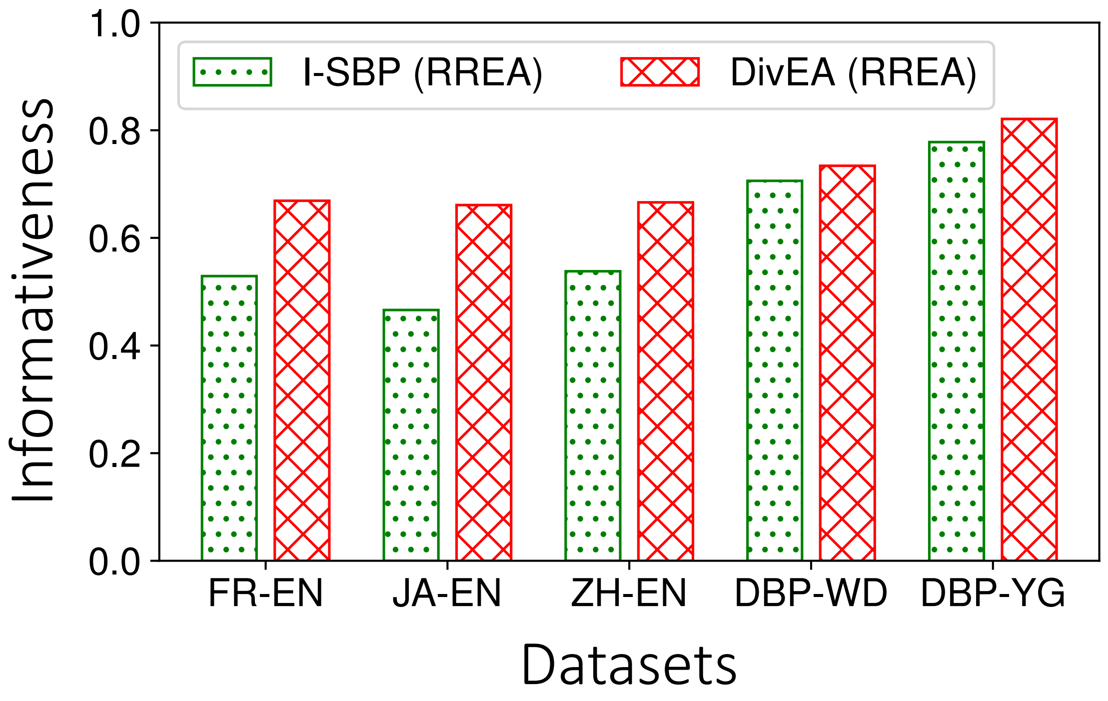
	15K			100K		2M
Method	FR-EN	JA-EN	ZH-EN	DBP-WD	DBP-YG	FB-DBP
CPS	0.817	0.718	0.826	0.542	0.486	0.237
DivEA	**0.881**	**0.892**	**0.880**	**0.830**	**0.893**	**0.507**
SBP	0.930	0.942	0.943	0.819	0.824	0.426
I-SBP	0.960	0.957	0.960	0.947	0.982	0.502
DivEA	**0.978**	**0.979**	**0.970**	**0.954**	**0.994**	**0.684**

Larger subtask size

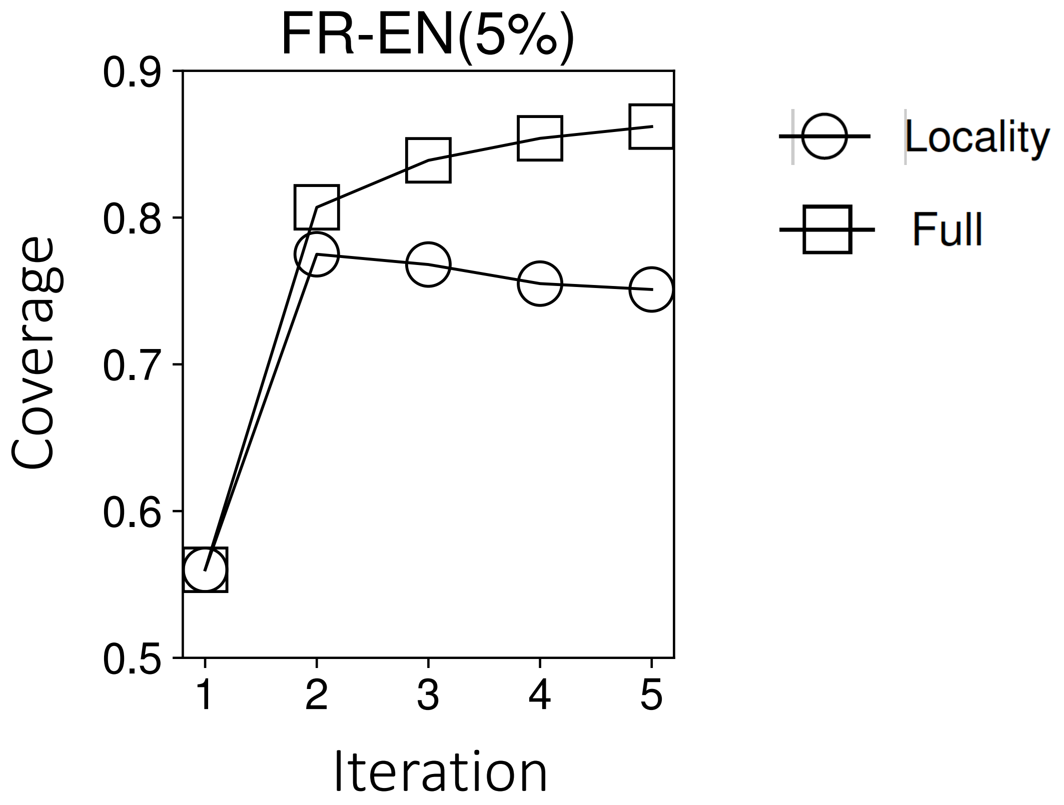
Results: Comparison with Baselines

- Informativeness of context graphs
  - Metric: percentage of found mappings over all mappings contained by the subtasks.





Results: Progressive Process of Counterpart Discovery



- Locality-based weight leads to decent performance
- EA model boosts it further

Conclusion

DivEA: a high-quality task division framework for large-scale EA.

- Dividing unmatched source entities + counterpart discovery + building context graphs.
- Progressive process.
- Building and running subtasks independently (for parallellization).

Code & data: https://github.com/uqbingliu/DivEA

Acknowledgement

Thank you for listening!

bing.liu@uq.edu.au
https://uqbingliu.github.io/
@BingLiu1011

SIGIR