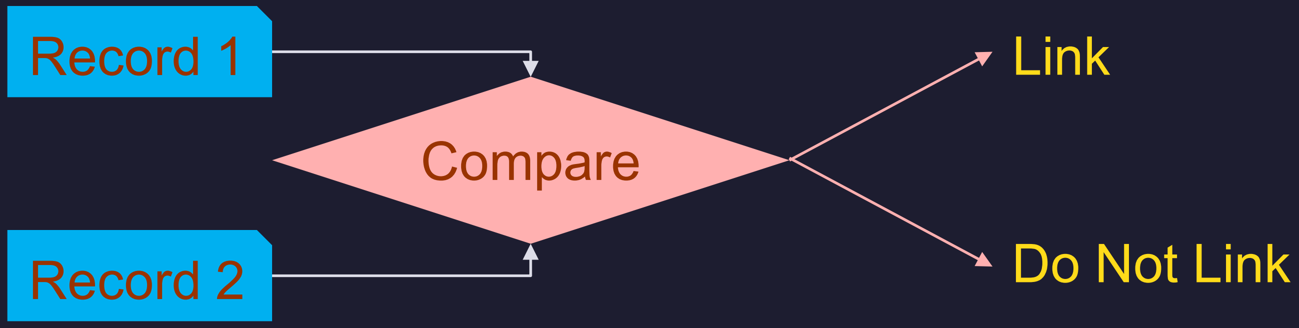
Supported in part by NSF grant IIS 1741022 and 1934565

Scaling Record Linkage

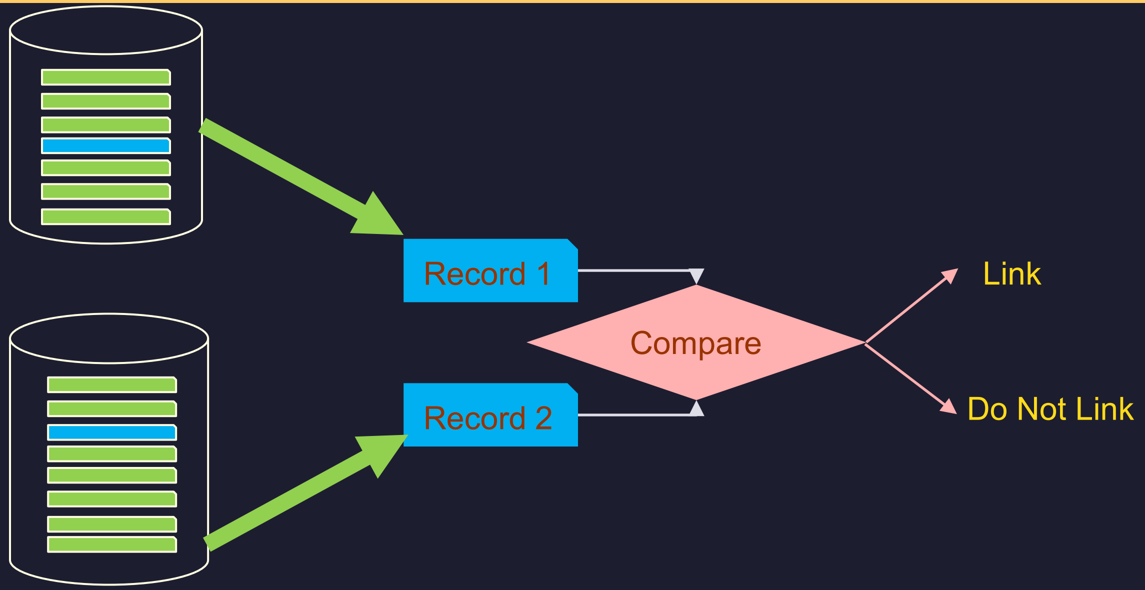
H. V. Jagadish
University of Michigan

Core Operation



- Compare operation can consider more than just the records, and can be very sophisticated (e.g. use AI methods).
- So, can be computationally expensive.

Repeated Core Operation

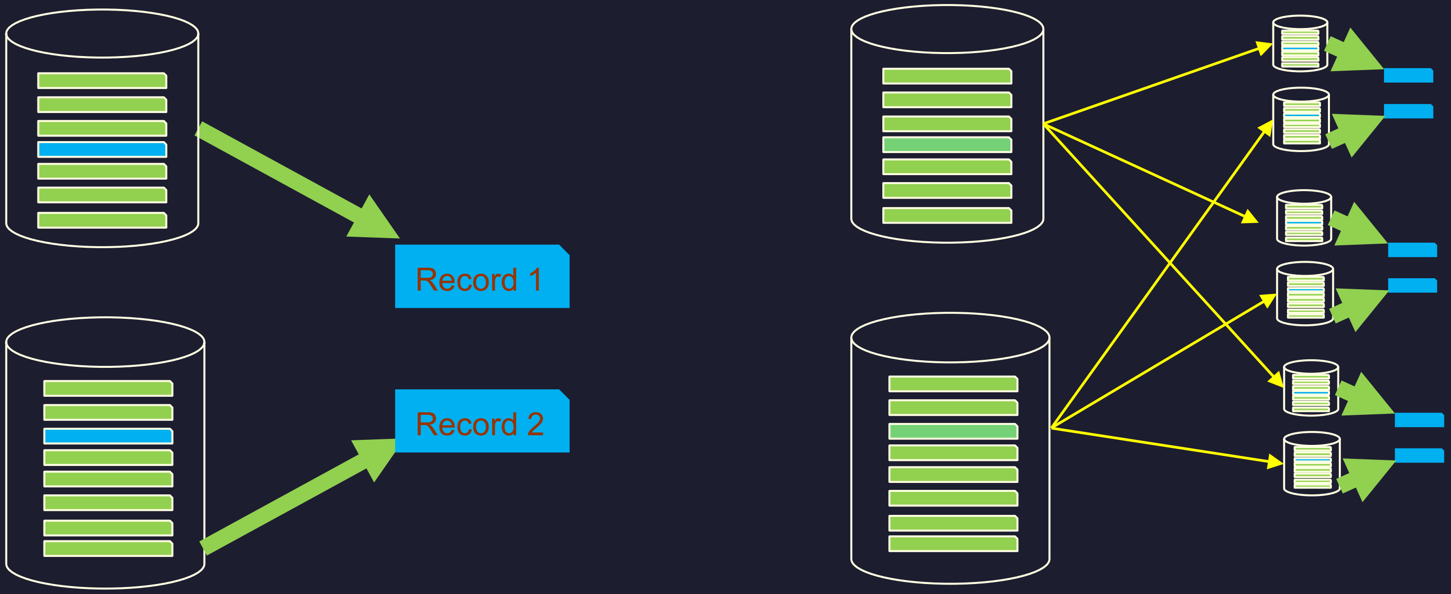


- Compare operation can be computationally expensive.
- Performed once for each pair of records.
- That is, *product* of record set sizes.
  - With 1,000 records in each record set, requires 1 million compare operations.
  - With 1 million records in each record set, requires 1 trillion compare operations.

Apply More Resources

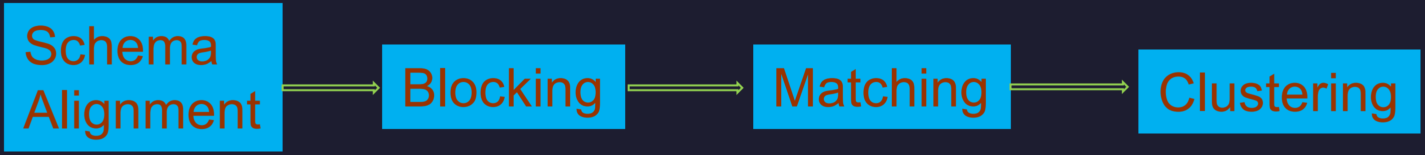
- Limits to processing power of any one machine.
- Compare tasks can be performed in parallel.
- Be smart about distributing tasks across a bunch of processors. E.g.
  - Evenly distribute load
  - Minimize data transfer

Reduce Number of Compare



- Eliminate pairs that couldn’t possibly match.
  - At least, unlikely to match.
- E.g. Use last name field as a basis to divide each record set into ”blocks”, and perform pairwise comparison only for records from corresponding blocks.
- With 1 million records in each record set, with blocks of 10, requires only 10 million compare operations.

Typical Workflow



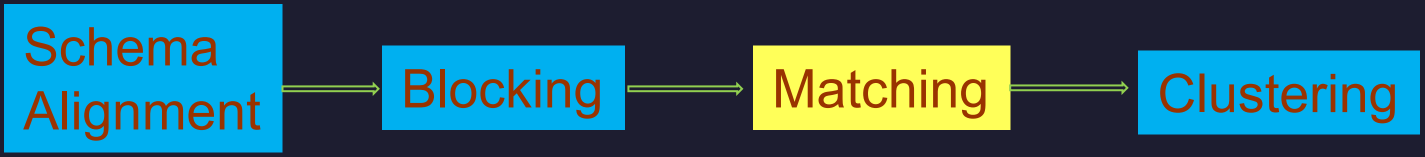
Schema Alignment

Blocking

Matching

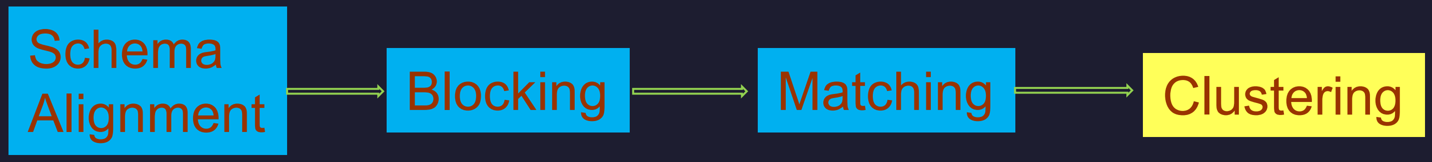
Clustering

Typical Workflow



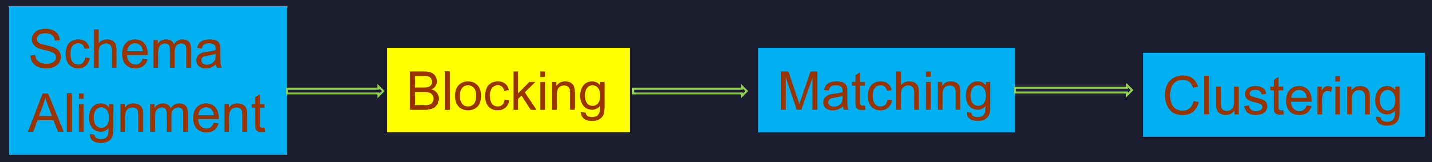
- Central operation
- Typically expensive
- Typically pairwise

Typical Workflow



- Multi-source linkage
- E.g. on the web

Typical Workflow



- Divide record set into blocks.
- Must be performed cheaply.
- Based only on an individual record
  - Without comparing with others
- E.g. Use a hash to partition.
- E.g. Last name + Zipcode

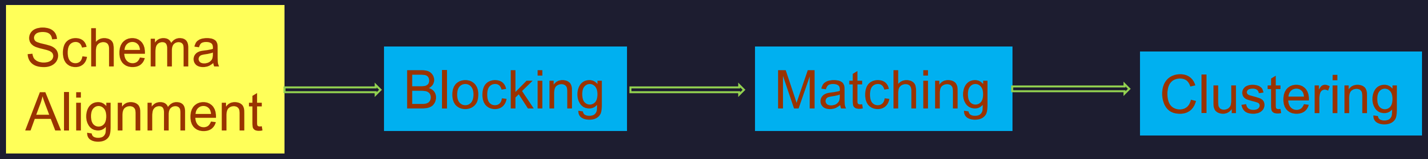
Data often has errors

- E.g. Misspelt last name
- E.g. Typographical errors
- Simple blocking can put related records in different blocks, and this is not recoverable.
- Fix by having larger (potentially overlapping blocks).
  - E.g. consider letter n-grams

Limits to Basic Blocking

- Need to identify “must have” conditions.
- E.g. Changed Last Name cannot be handled.

Typical Workflow



Schema Alignment

Blocking

Matching

Clustering

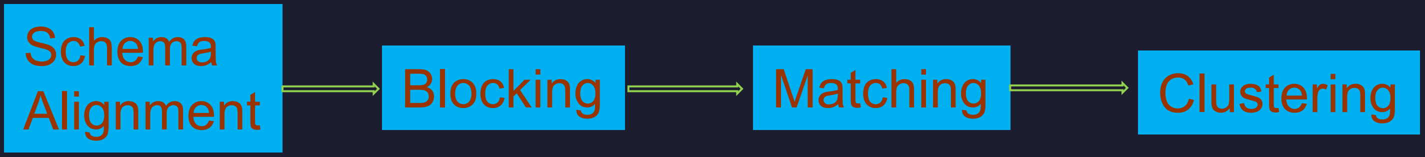
- Sources typically independent
  - Except for duplicate merging
- Corresponding attributes may have different names.
  - May even be differently structured
    - Name vs. firstname, lastname
  - Or differently expressed
    - Date formats
    - Units of measurement

Heterogeneity

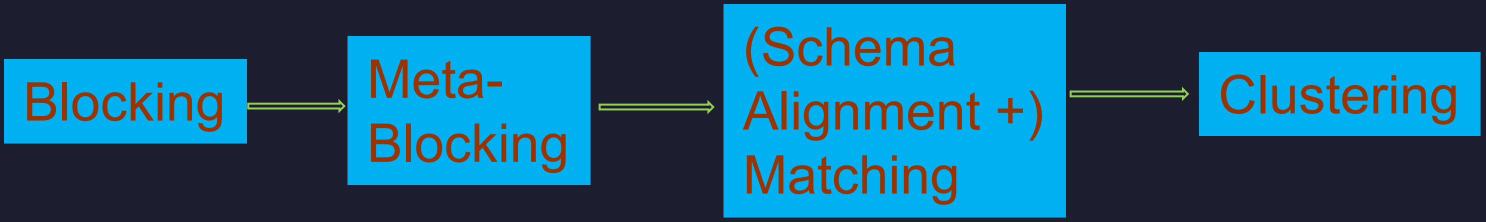
- Schema alignment is hard
- Often imperfect
- Would rather address (some of it) at match time

- There may even be no schema for some records, e.g. in NL text.

Modified Workflow



Schema Alignment → Blocking → Matching → Clustering



Blocking → Meta-Blocking → (Schema Alignment +) Matching → Clustering

Meta-blocking

- Create lots of (overlapping) blocks.
- Be generous in block creation and record assignment.
- Each record assigned to multiple blocks.

- Use meta-blocking to clean this up.

Meta-Blocking Example

p1

Name: John Abram Jr
profession: car seller
year: 1985
Addr.: Main street

p3

name1: Jon Jr
name2: Abram
birth year: 85
job: car retail
Loc: Main st.

p2

FirstName: Ellen
SecondName: Smith
year: 85
occupation: retail
mail: Abram st. 30 NY

p4

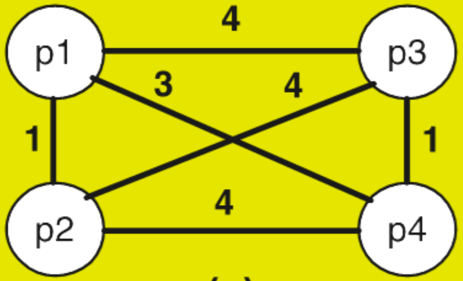
full name: Ellen Smith
b. date: May 10 1985
work info: retailer
loc: Abram street NY

(a)

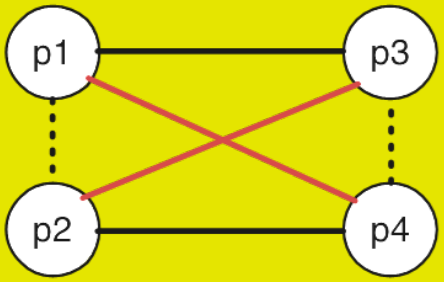
Ellen	**NY**	**Main**	**1985**	**car**	**Smith**
p2 p4	p2 p4	p1 p3	p1 p4	p1 p3	p2 p4

street	**Jr**	**retail**	**st**	**85**	**Abram**
p1 p4	p1 p3	p2 p3	p2 p3	p2 p3	p1 p2 p3 p4

(b)

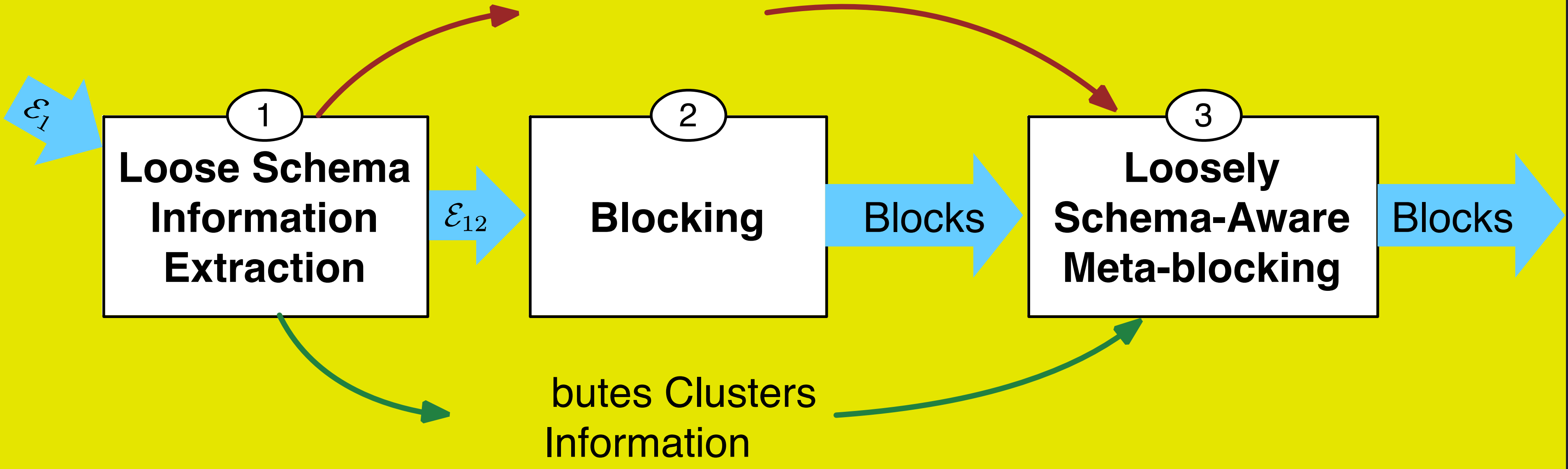


(c)

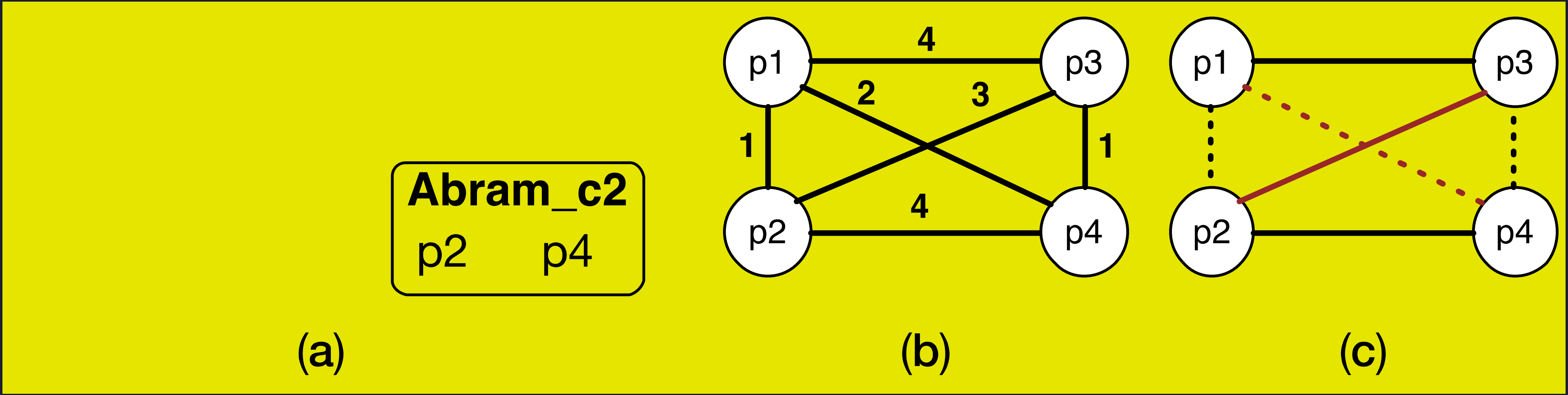


(d)

BLAST

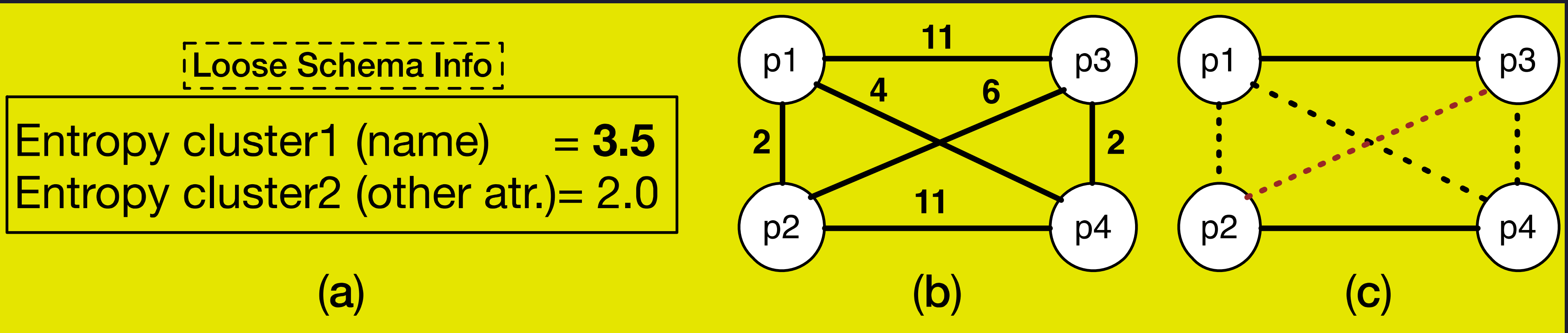

$\mathcal{E}_1$
1
Loose Schema Information Extraction
$\mathcal{E}_{12}$
2
Blocking
Blocks
3
Loosely Schema-Aware Meta-blocking
Blocks
butes Clusters
Information

Attribute Clustering



- Cluster attributes very roughly into groups.
  - – Much easier than full schema alignment
- Block only for shared token in same group.

Attribute (Cluster) Entropy



Loose Schema Info

Entropy cluster1 (name)   = **3.5**
Entropy cluster2 (other atr.)= 2.0

(a)

- Not all attributes are equally informative
- Compute attribute entropies
  - At attribute cluster level
- Weight edges by entropy for meta-blocking

Conclusion

- Record linkage is messy.
- Many clever methods to match (not discussed today).
- But can quickly get expensive.
- Use blocking, and meta-blocking to scale.
- Also use parallelism
  - We have work to parallelize meta-blocking