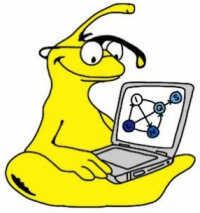
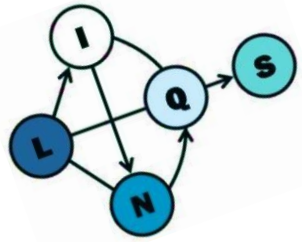
Entity Resolution

Eriq Augustine and Golnoosh Farnadi
UC Santa Cruz
MLTrain 2018

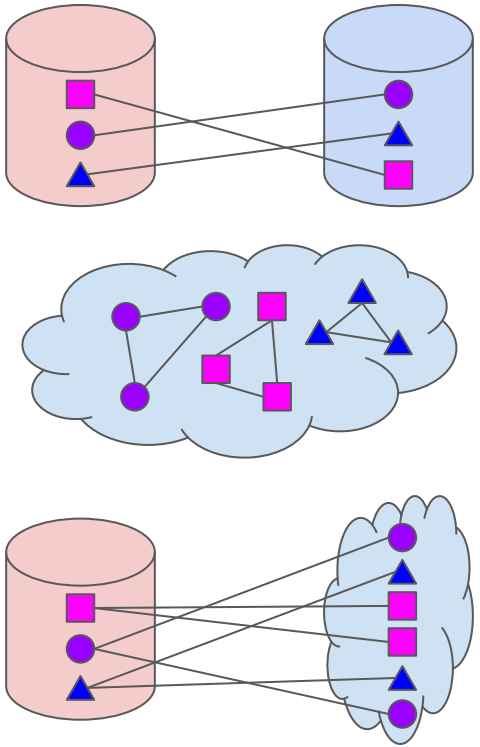


psl.linqs.org
github.com/linqs/psl



What is Entity Resolution?

- Entity Resolution comes in several variants:
  - Record Linkage
    - Matching between two (mostly) deduplicated data sources
    - Makes the 1-1 assumption
  - Deduplication
    - Given a single collection of references, find all references that refer to the same entity.
  - Reference Matching
    - Given a deduplicated and a noisy source, match all the noisy references to the deduplicated entities.



Getting the Code

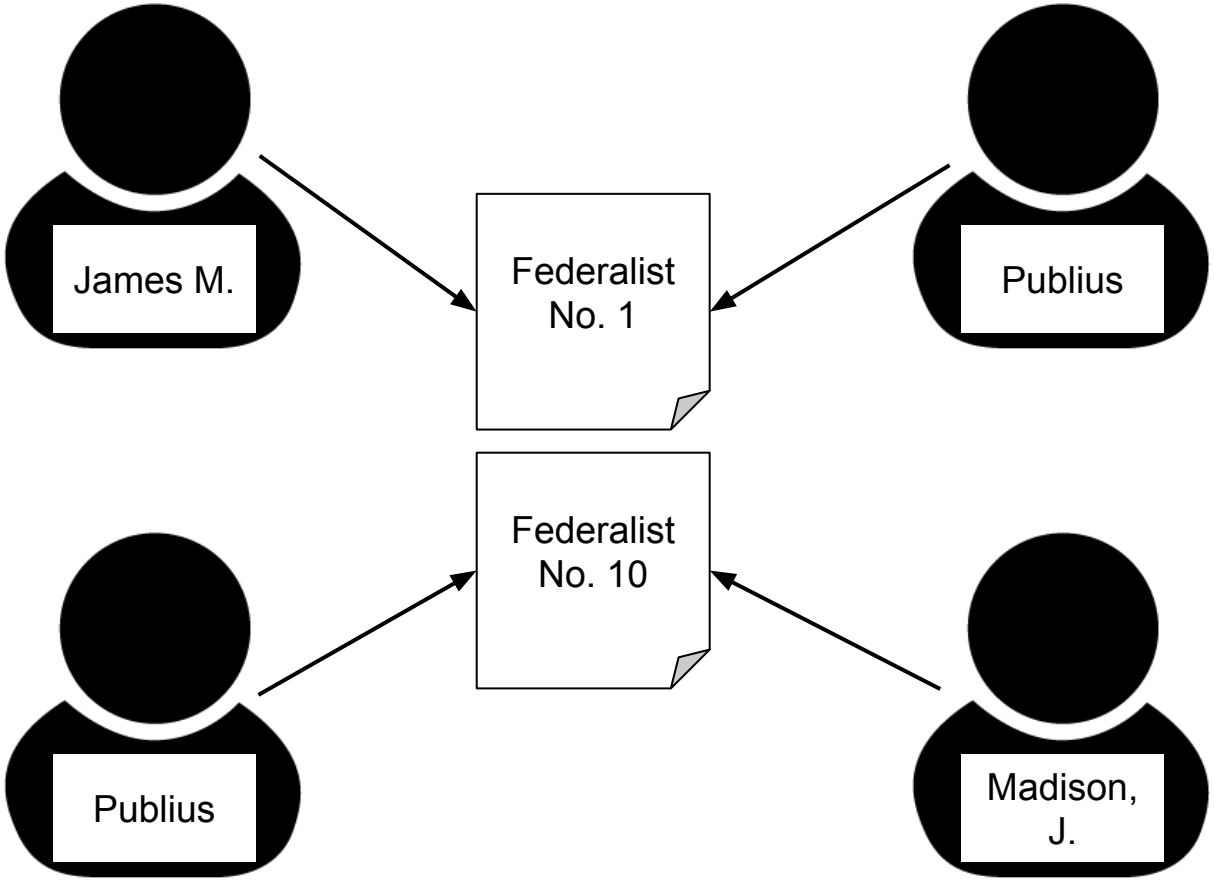
```
git clone https://github.com/linqs/psl-examples.git

cd psl-examples/entity-resolution/cli

git checkout uai18

./run.sh
```

All examples available at: https://github.com/linqs/psl-examples

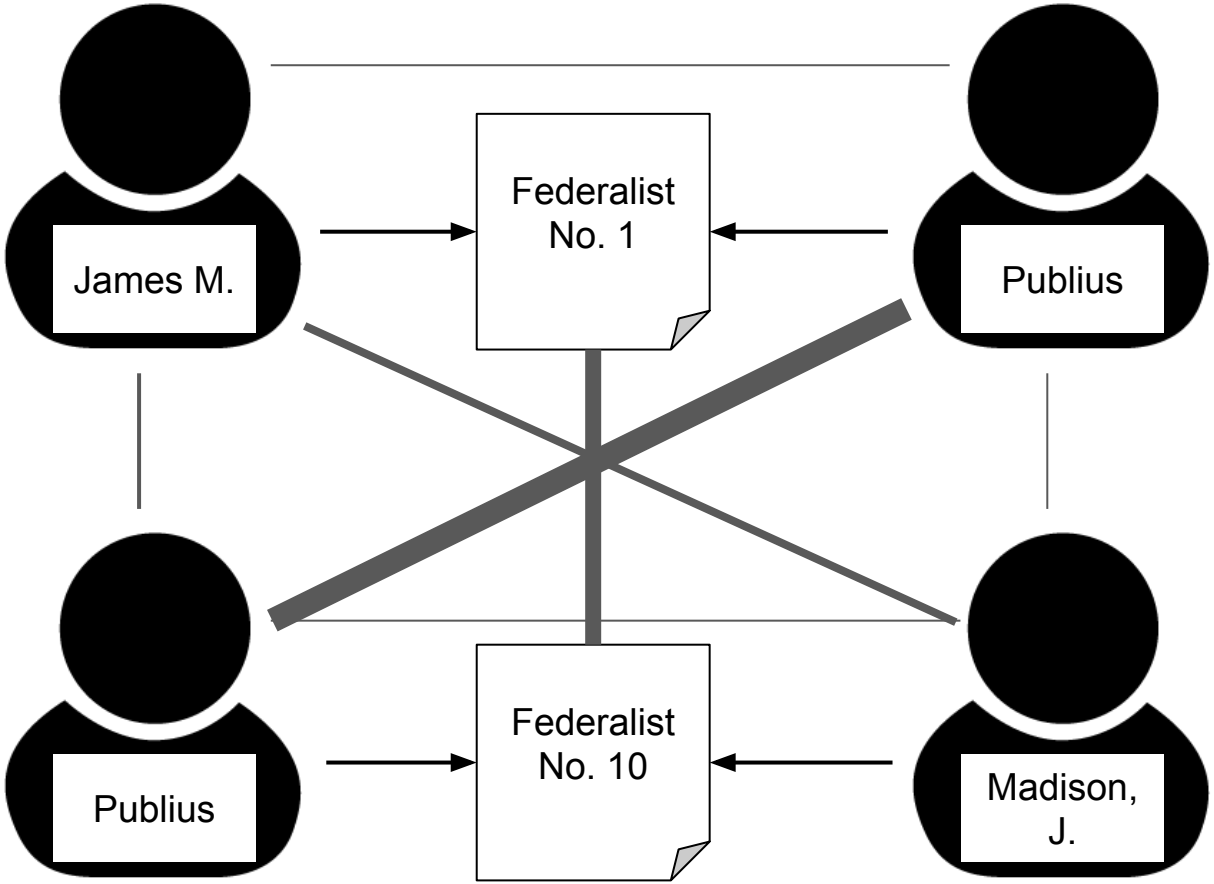
Data

- Citation Network
- Deduplicate
  - Authors
  - Papers
- CiteSeer

Size	Authors	Papers
Small	1136	864
Medium	**1813**	**1143**
Large	2892	1504

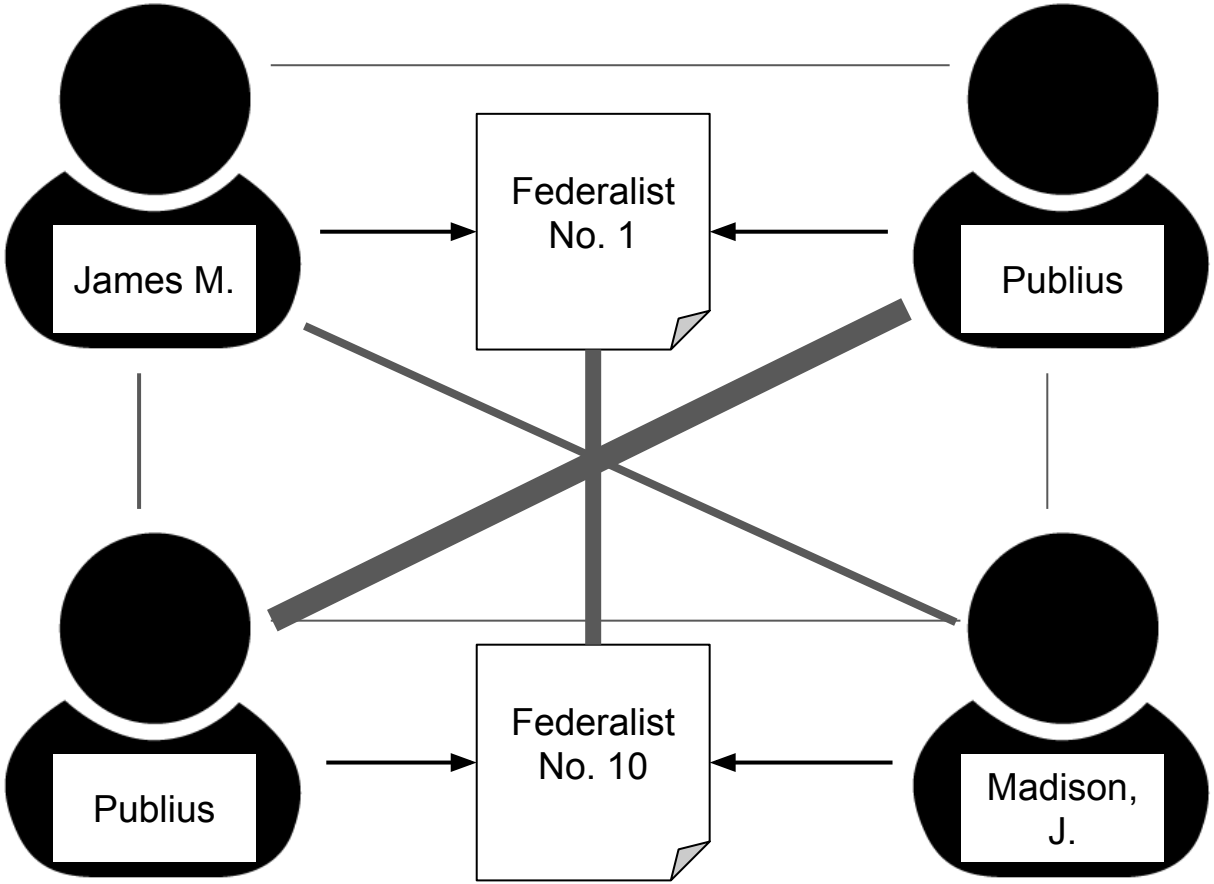


Initial Model



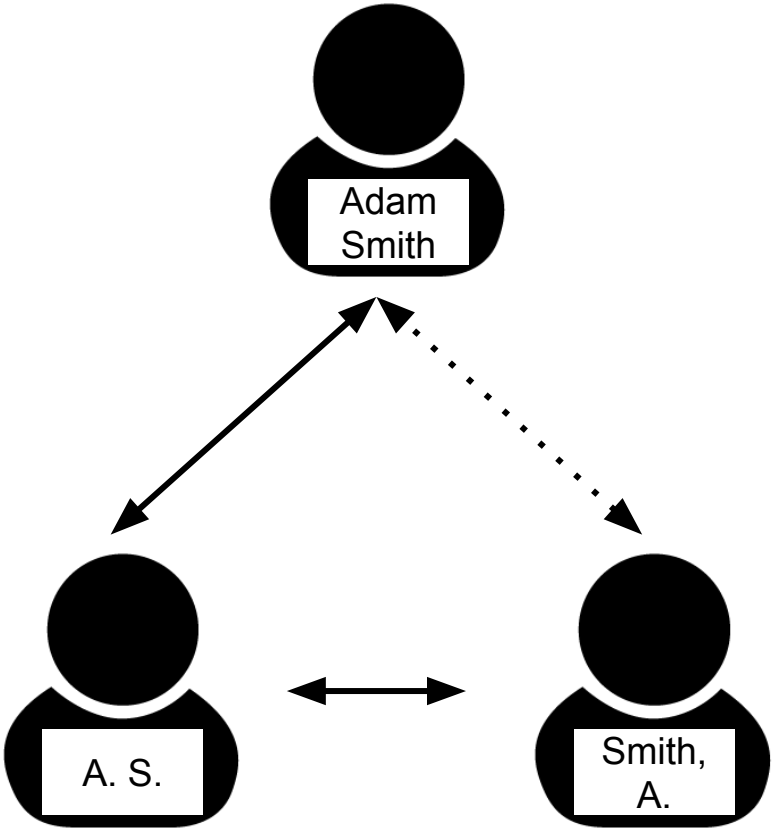
Initial Model

```
AuthorName(A1, N1)
     & AuthorName(A2, N2)
     & SimName(N1, N2)
     -> SameAuthor(A1, A2)


PaperTitle(P1, T1)
     & PaperTitle(P2, T2)
     & SimTitle(T1, T2)
     -> SamePaper(P1, P2)
```



Transitive Equality

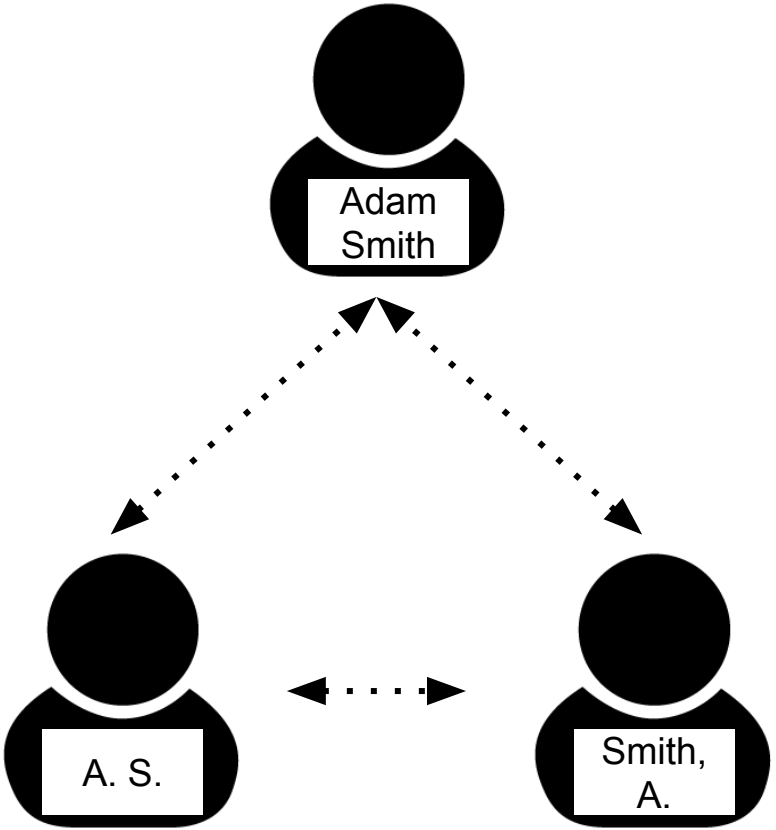
Exploit relational nature of similarity in ER.



Transitive Equality

Exploit relational nature of similarity in ER.

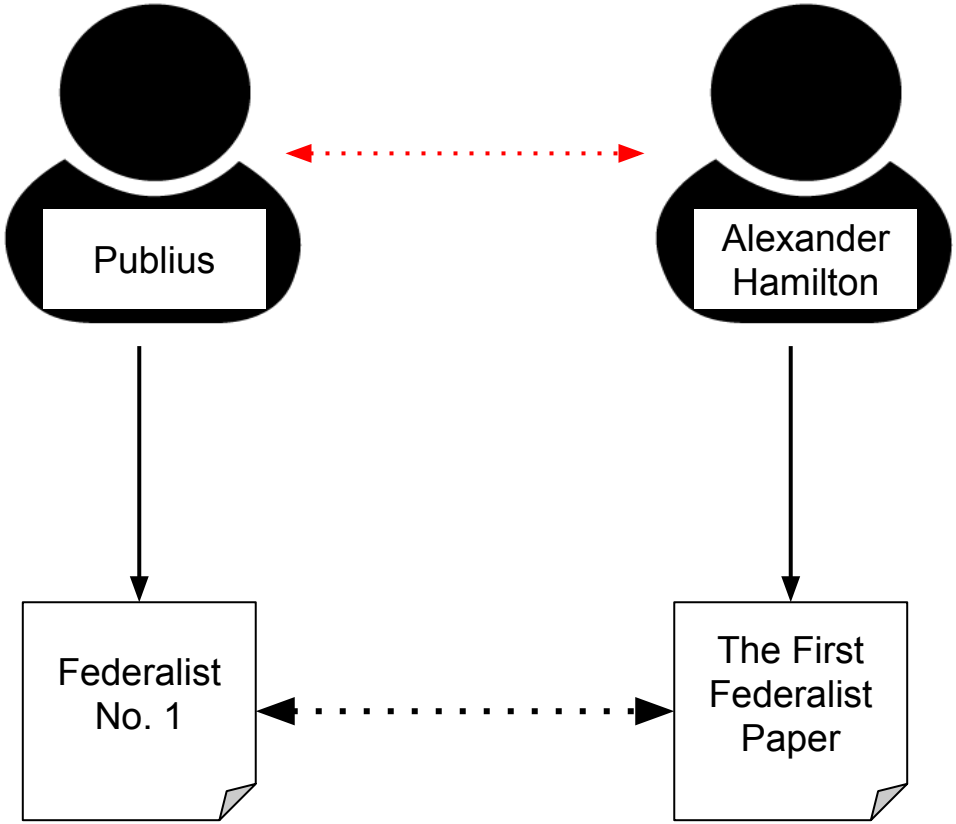
```
SameAuthor(A1, A2)
     & SameAuthor(A2, A3)
     -> SameAuthor(A1, A3)
```



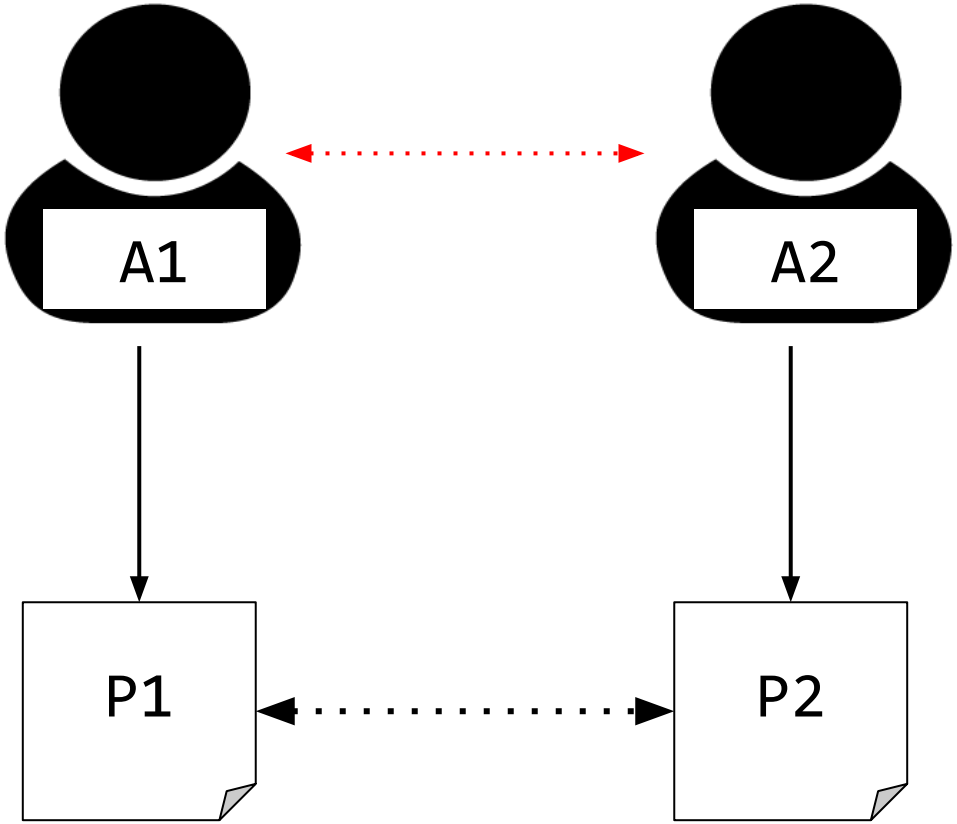
Transitive Relational

What other transitive relational rules can we get?

Transitive Relational



Transitive Relational

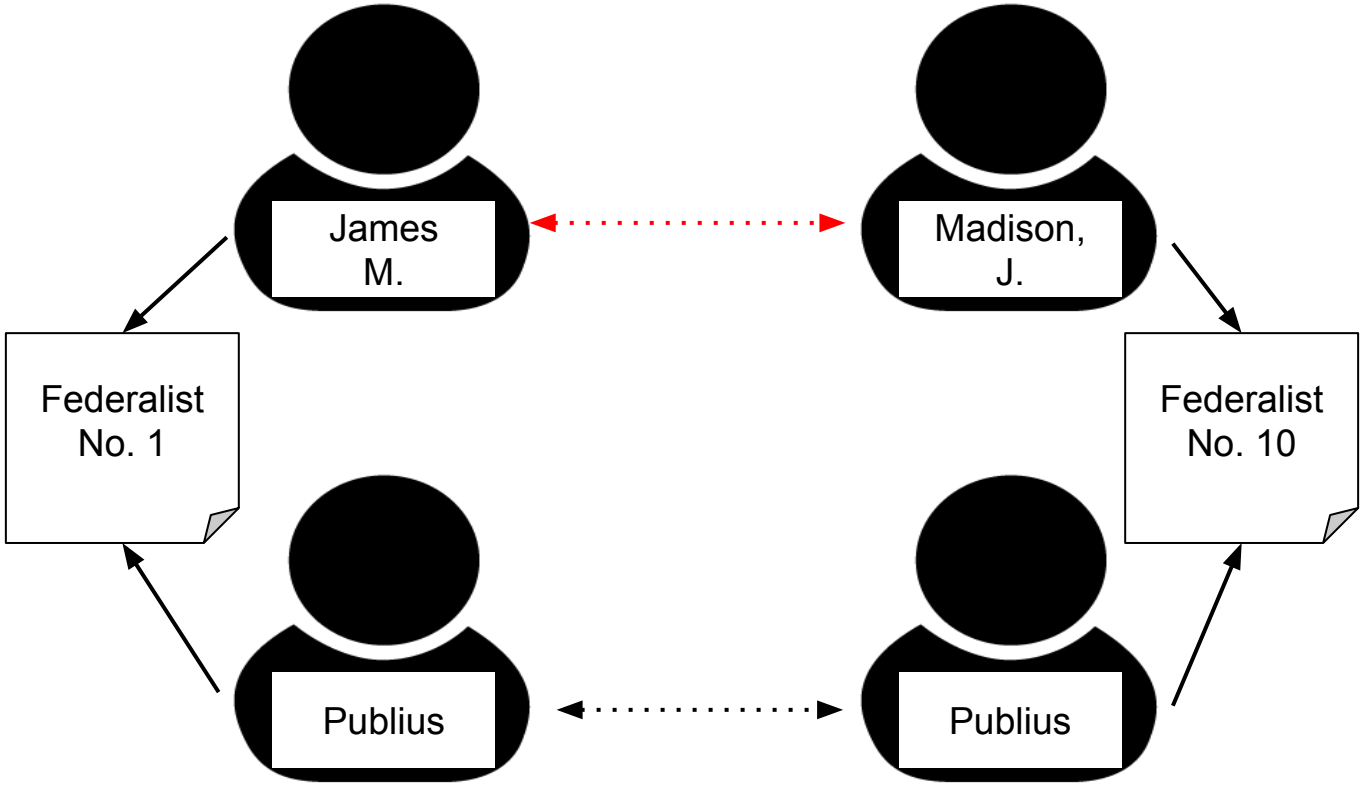



```
AuthorOf(A1, P1) & AuthorOf(A2, P2) & SamePaper(P1, P2) -> SameAuthor(A1, A2)
```

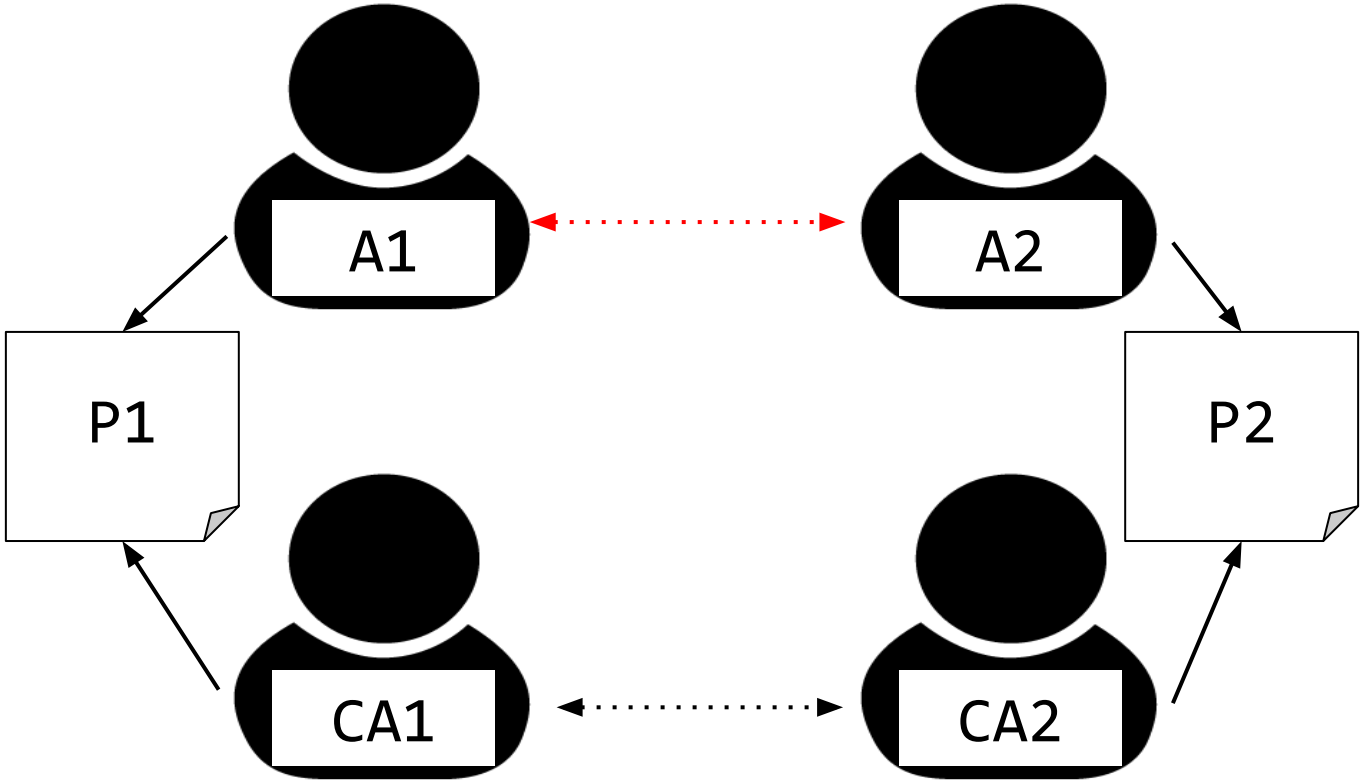
Transitive Relational

What other transitive relational rules can we get?

Transitive Relational



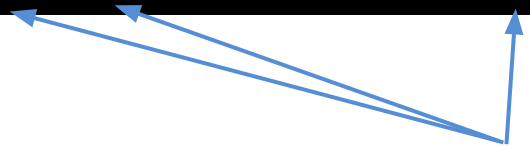
Transitive Relational




```
AuthorOf(A1, P1) & AuthorOf(A2, P2) & AuthorOf(CA1, P1) & AuthorOf(CA2, P2)
    & SameAuthor(CA1, CA2) -> SameAuthor(A1, A2)
```

Transitive Blowup!

```
SameAuthor(A1, A2) & SameAuthor(A2, A3) -> SameAuthor(A1, A3)
```



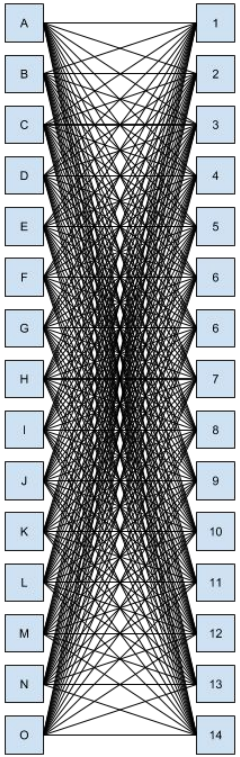
Arbitrarily choose **three** authors.

Recall we have 1813 authors.

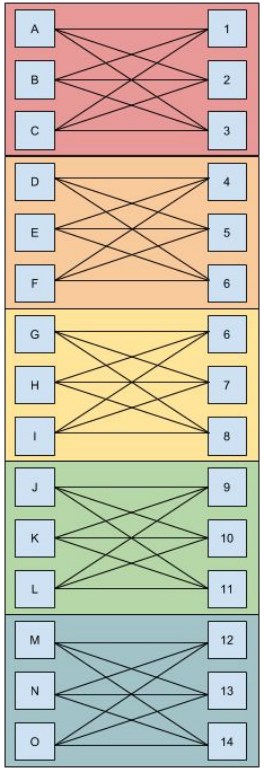
$$\binom{1813}{3}$$

~ 1 Billion Ground Rules

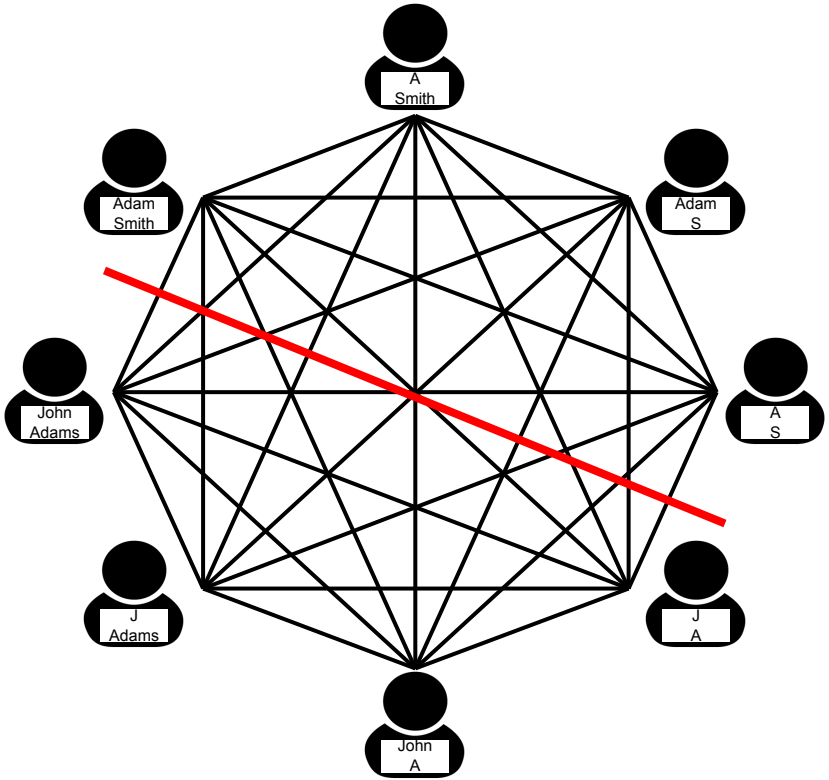
Blocking



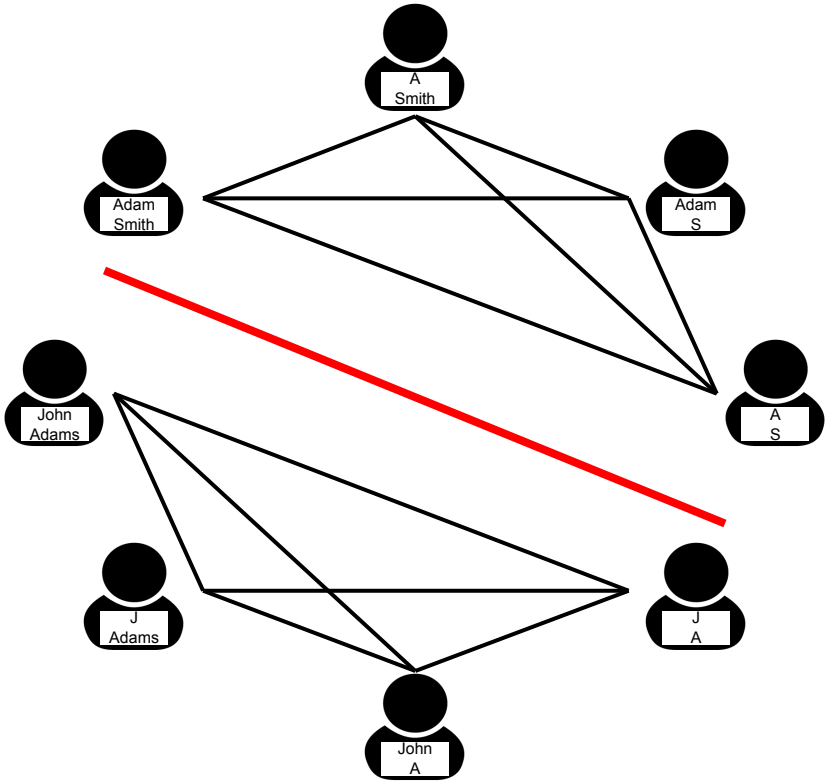
- Blocking is reducing the number of ground potentials using some computed heuristic(s).
- In PSL, this is done by adding predicates that induce sparsity in the MRF.



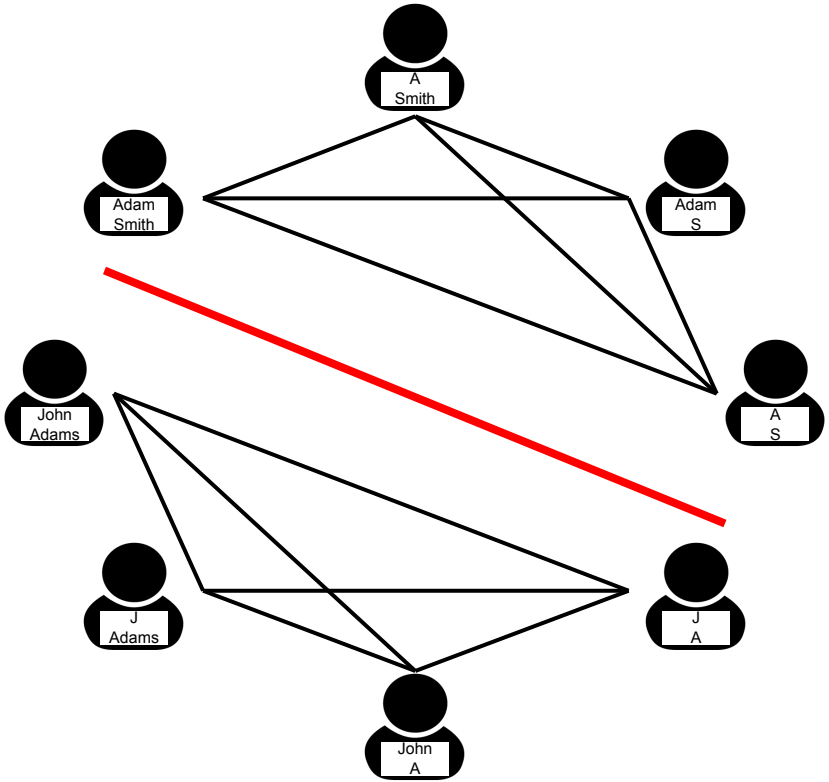
Blocking



Blocking



Blocking




```
AuthorBlock(A1, B) & AuthorBlock(A2, B) & AuthorBlock(A3, B)
    & SameAuthor(A1, A2) & SameAuthor(A2, A3) -> SameAuthor(A1, A3)
```

Blocking

```
AuthorBlock(A1, B) & AuthorBlock(A2, B) & AuthorBlock(A3, B)
    & SameAuthor(A1, A2) & SameAuthor(A2, A3) -> SameAuthor(A1, A3)


AuthorBlock(A1, B1) & AuthorBlock(A2, B1)
    & AuthorBlock(CA1, B2) & AuthorBlock(CA2, B2)
    & AuthorOf(A1, P1) & AuthorOf(A2, P2)
    & AuthorOf(CA1, P1) & AuthorOf(CA2, P2) & SameAuthor(CA1, CA2)
    -> SameAuthor(A1, A2)


AuthorBlock(A1, B) & AuthorBlock(A2, B)
    & AuthorOf(A1, P1) & AuthorOf(A2, P2) & SamePaper(P1, P2) -> SameAuthor(A1, A2)
```

Blocking - How to Make Blocks

- How can we block authors?
- Need to tradeoff:
  - **Speed**
  - **Recall**
  - **Precision**

Blocking - How to Make Blocks

- How can we block authors?
- Need to tradeoff:
  - **Speed**
  - **Recall**
  - **Precision**
- Alphabetized Initials?

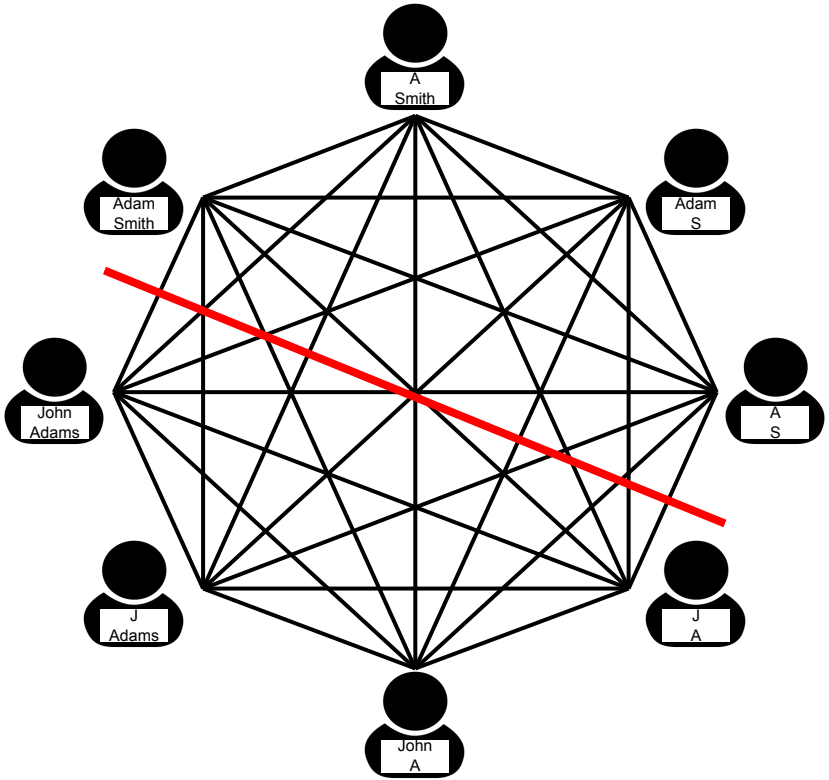
Pros:

- Fast
- Catch most misspellings
- Catch initials
- Catch Different Order
- Catch some nicknames

Cons:

- Miss some nicknames
- Miss totally different names

Blocking



Results - Quality

Size	Transitive Relational	Blocking?	Time (sec)	Author F1
Medium	None	No	166	0.7996
Medium	Equality	Yes	176	0.8157
Medium	Coauthor	Yes	173	0.8113
Medium	Paper	Yes	166	0.8158
Medium	All	Yes	180	**0.8467**

Results - Speed

Size	# Ground Rules	Transitive Relational	Blocking?	Time (sec)	Author F1
Small	220 M	Equality	No	21600+	N/A
Small	0.5 M	All	Yes	55	0.80946
Medium	1.5 M	All	Yes	180	0.846722
Large	3.3 M	All	Yes	413	0.734253