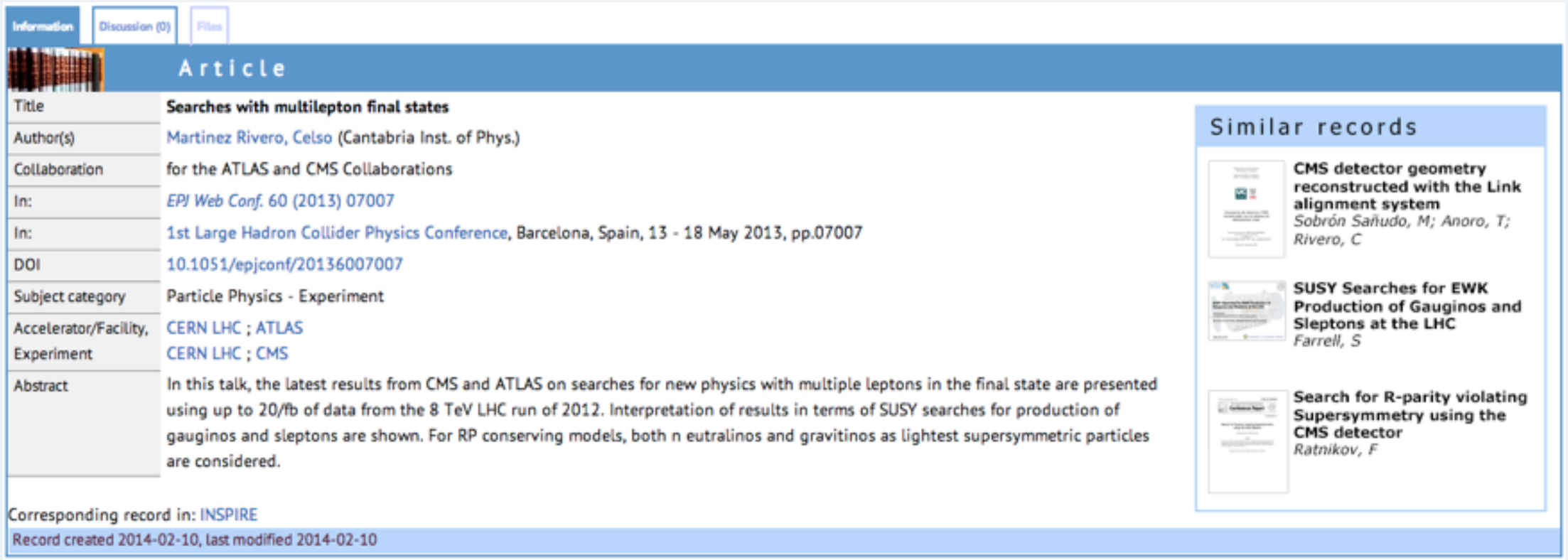
MLTDL - More Like This for Digital Libraries

CERN spin-off

TIND

What is More Like This?



CERN spin-off

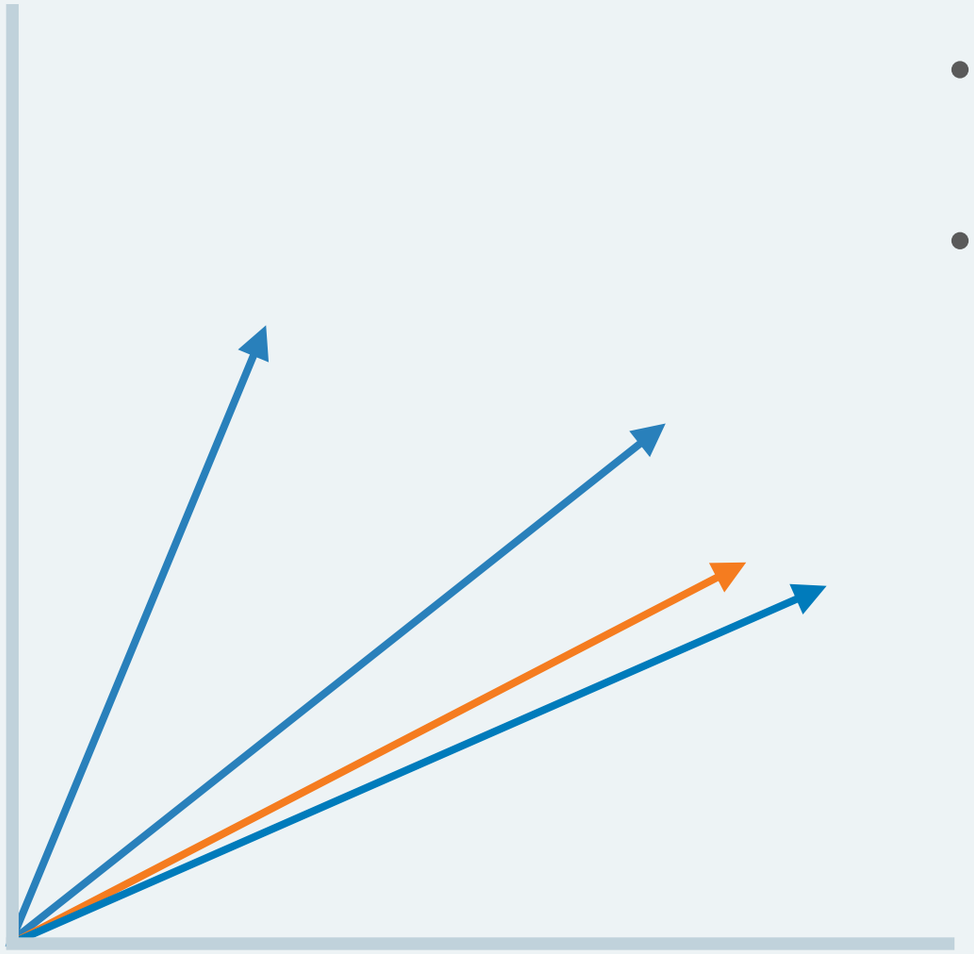
TIND

Why More Like This?

CERN spin-off

TIND

How does MLT work today?

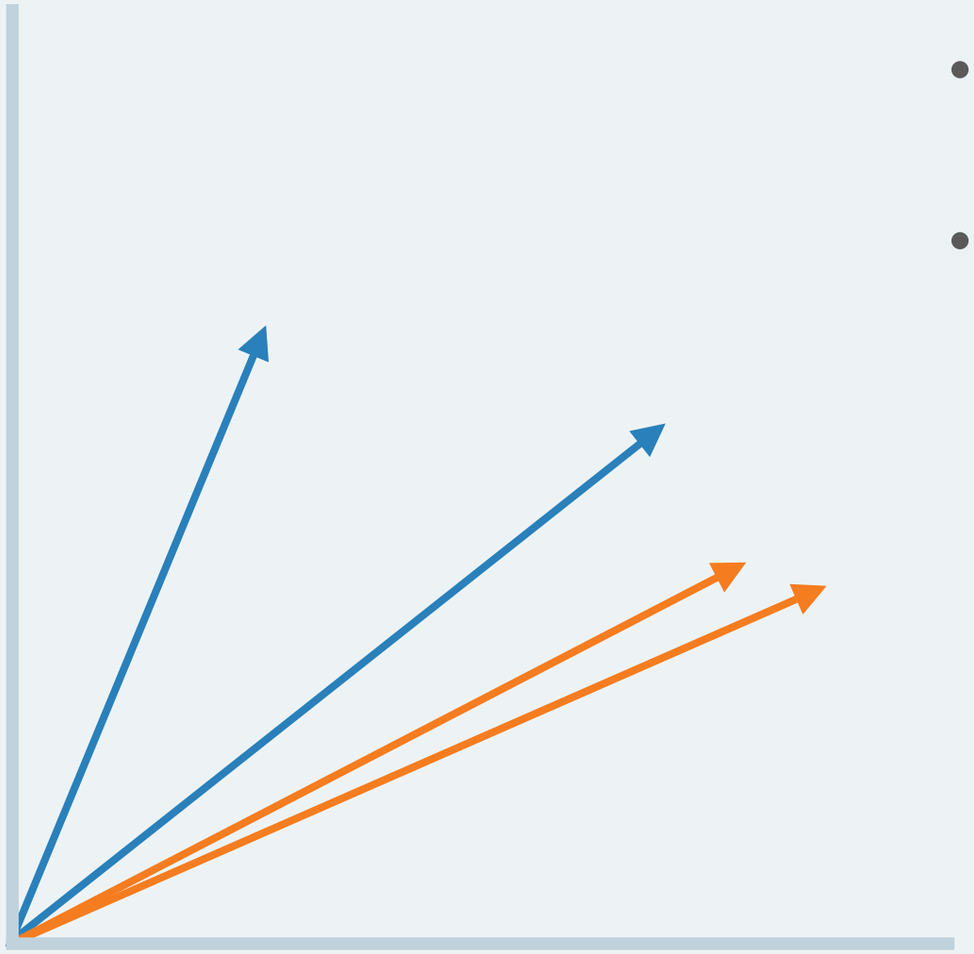


- TF-IDF
- Term Frequency
- Inverse Document Frequency

- Supported by popular indexers
  - Solr
  - ElasticSearch

spin-off

How does MLT work today?



- TF-IDF
- Term Frequency
- Inverse Document Frequency

- Supported by popular indexers
  - Solr
  - ElasticSearch

Can MLT do better?

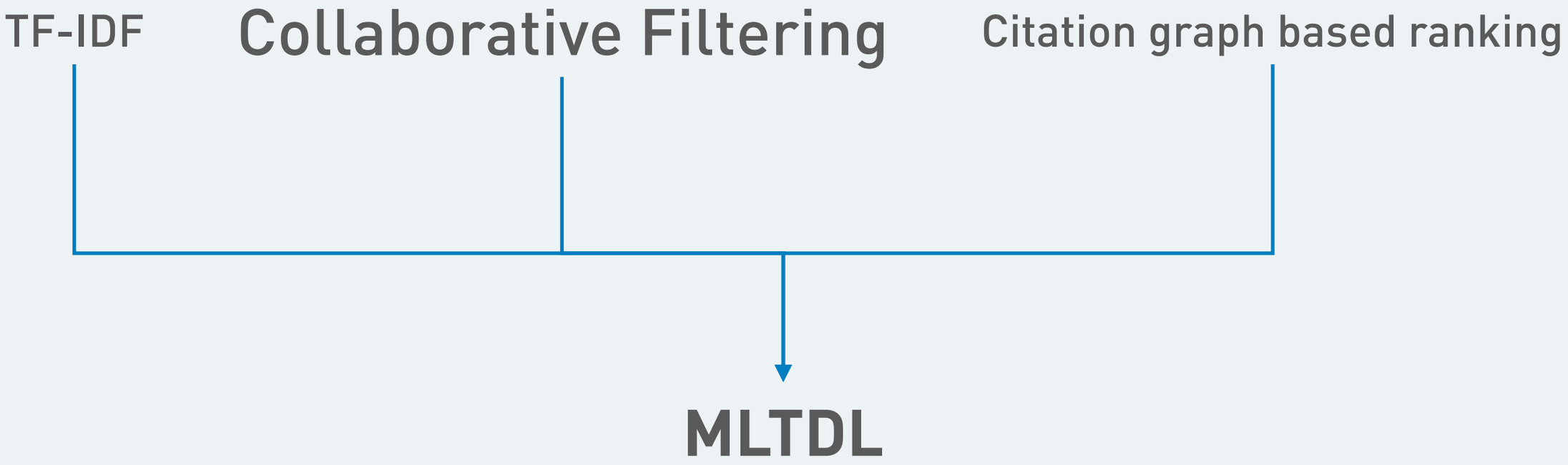
TF-IDF:

+ Efficient queries
+ Efficient indexing
+ Simple implementation

– Documents in the collection may be too similar
– No concept of user activity
– Uses only a small set of the information available

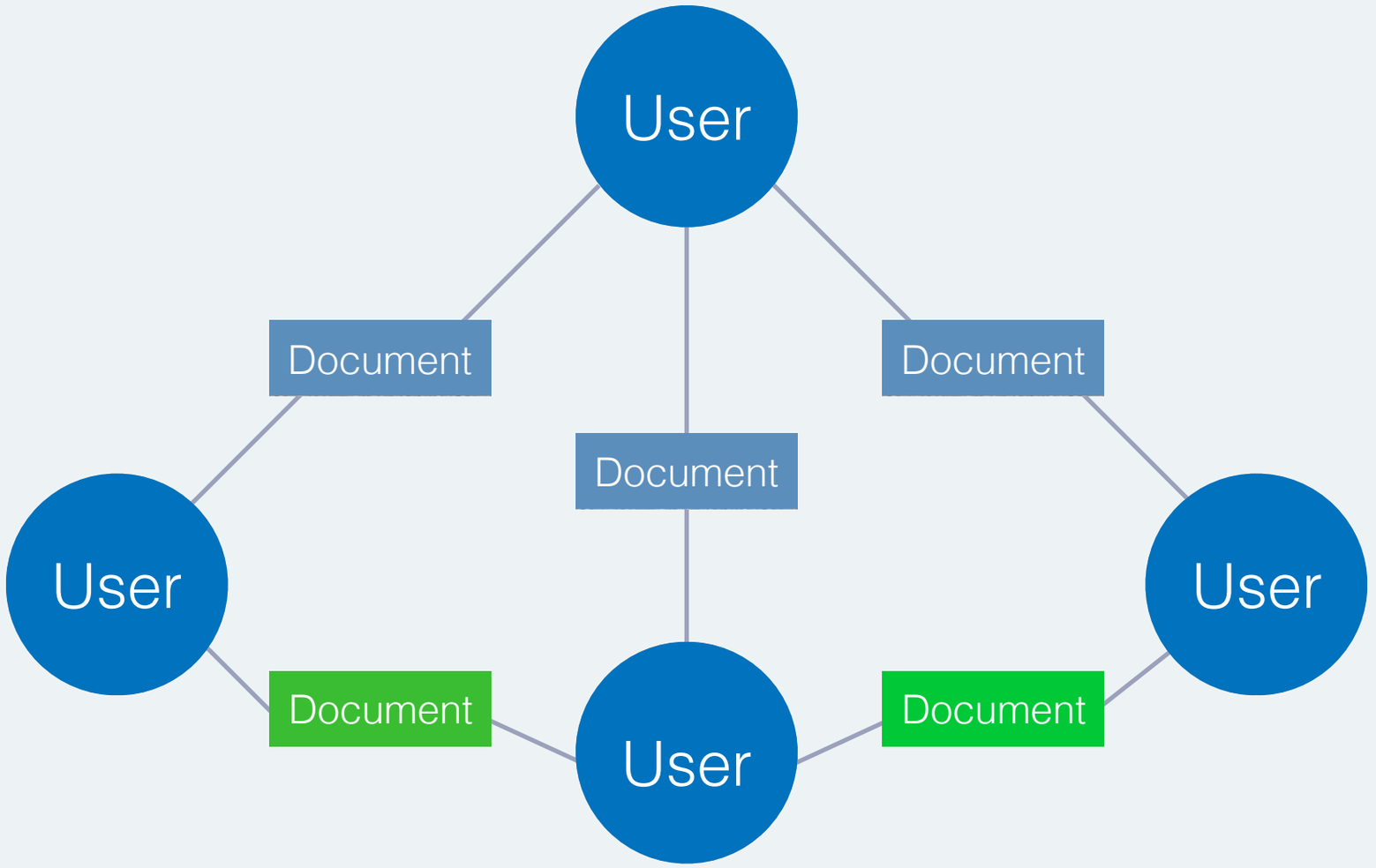
Introducing MLTDL

Utilizing the characteristic attributes of the digital library



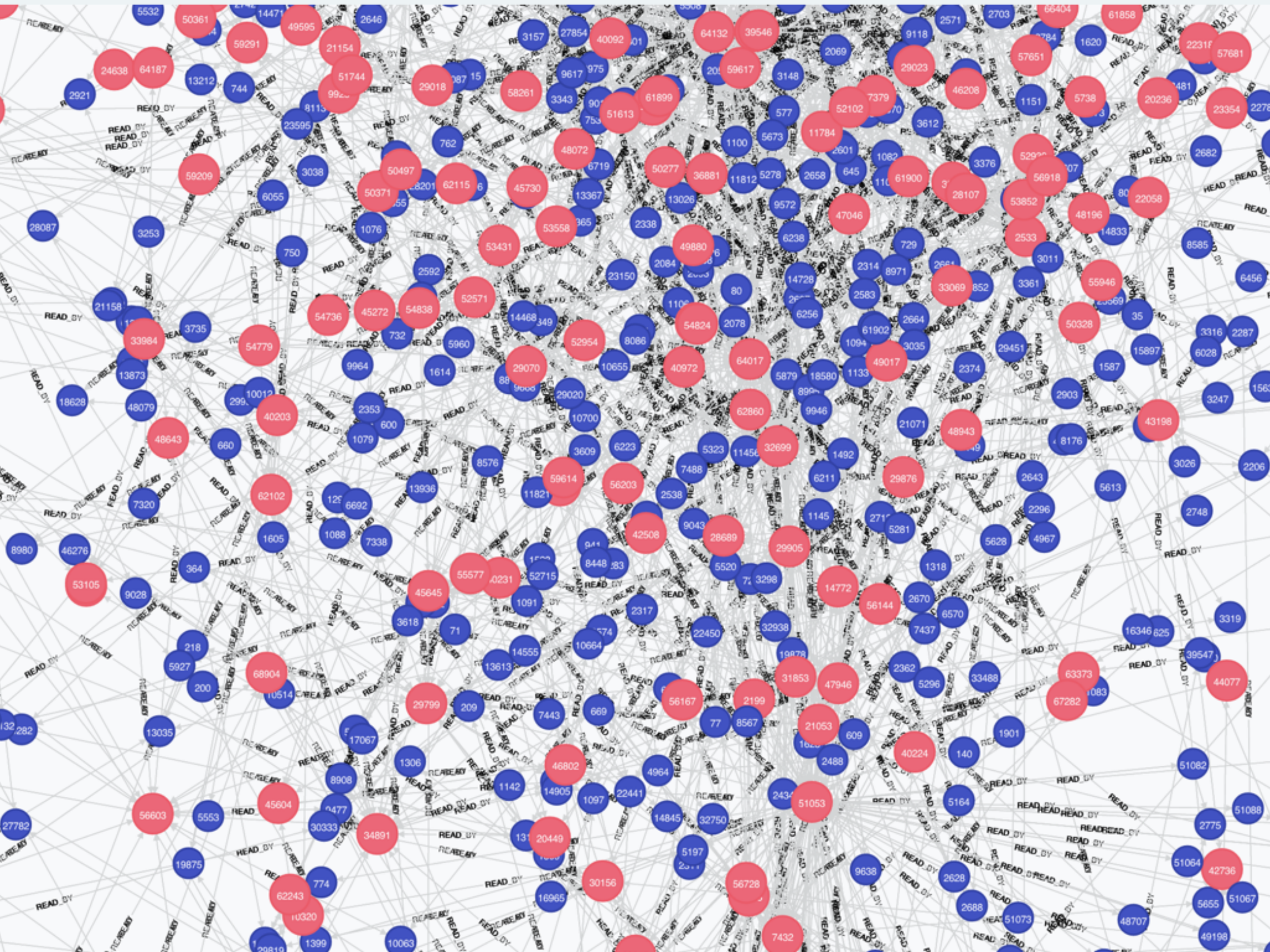
spin-off

Graph based collaborative filtering



spin-off

TIND


READ_BY
NEARREADY

Who are the users?

Bots

Overly curious

Librarians

Singleton

Developers

Real users

CERN spin-off

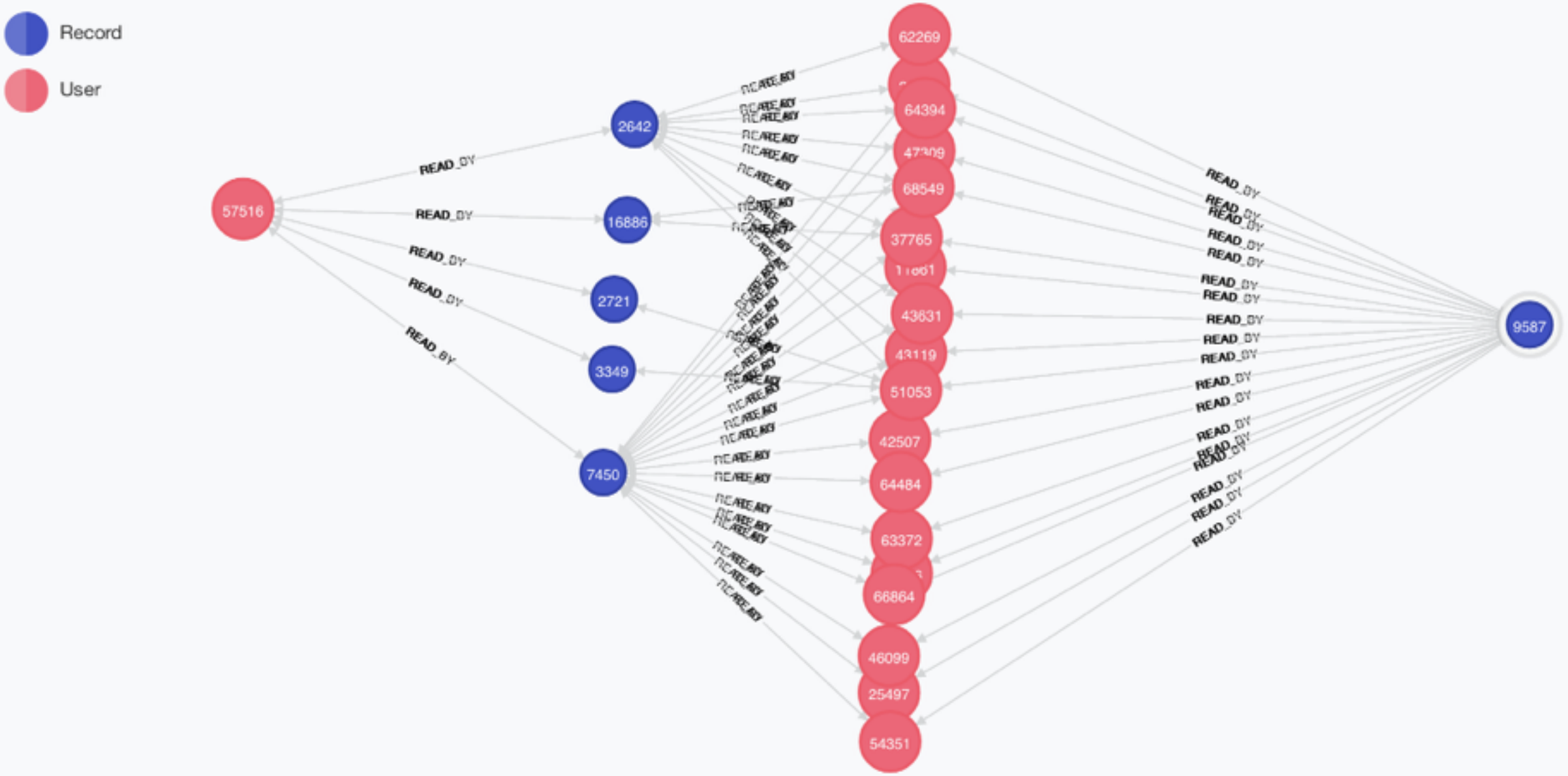
TIND

Doin' the diggin'

Automatic log analysis and graphing

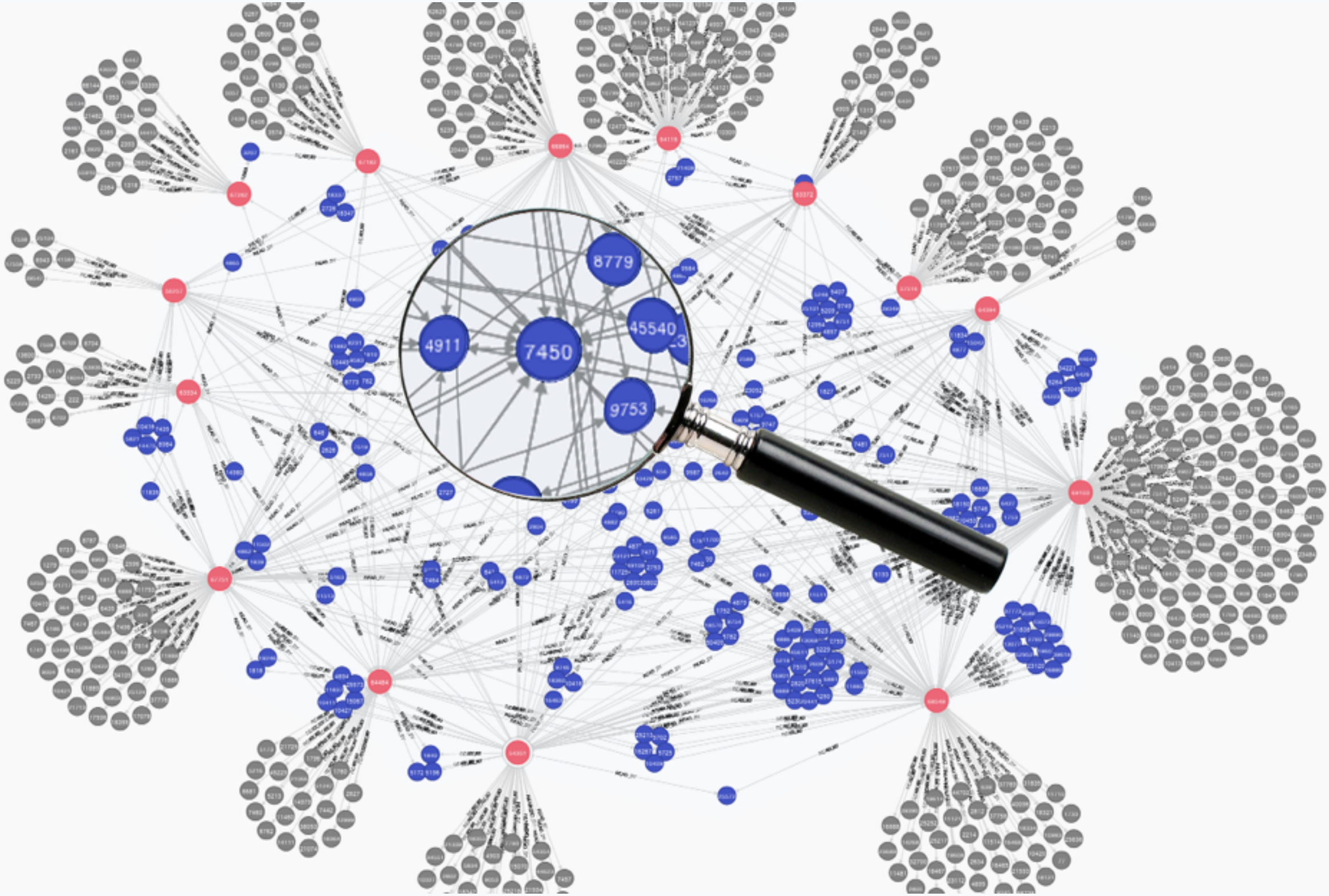
spin-off

TIND


Record
User
57516
READ_BY
2642
16886
2721
3349
7450
62269
64394
68549
37765
43631
51053
42507
64484
63372
66864
46099
25497
54351
9587

spin-off

Record
User



spin-off

TIND

Looking at record 1344131

From one users perspective

	More Like This	Our recommendation
1	1391730	1328960
2	1235963	1391730
3	1367755	1367755
4	789847	1378511
5	1399527	
6	1319540	
7	1378511	
8	1328960	

Record 2, 4 and 6 are not present in graph
Record 5 is not relevant according to our graph

spin-off

Looking at record 1344131

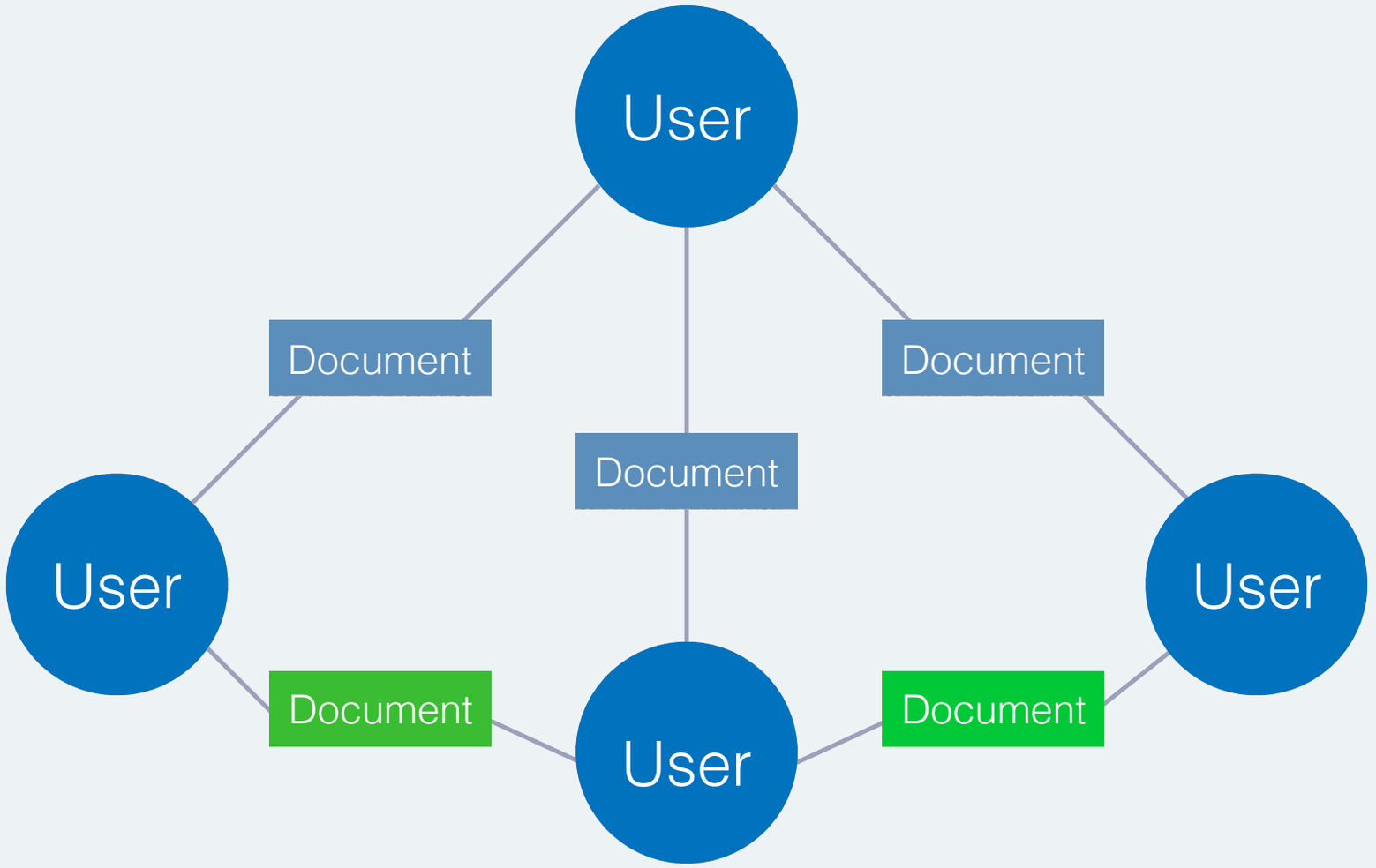
From another users perspective

	More Like This	Our recommendation
1	1391730	1367755
2	1235963	1399527
3	1367755	1378511
4	789847	
5	1399527	
6	1319540	
7	1378511	
8	1328960	

Record 2, 4 and 6 are not present in graph
Record 1 has been read by the user

spin-off

Questions?



spin-off