Project presentation

Machine Learning and Big Data Processing(ELEC–Y591)

Cédric Hannotier Mathieu Petitjean Hasan Can Yildirim

June 13, 2018

Outline

State of the art

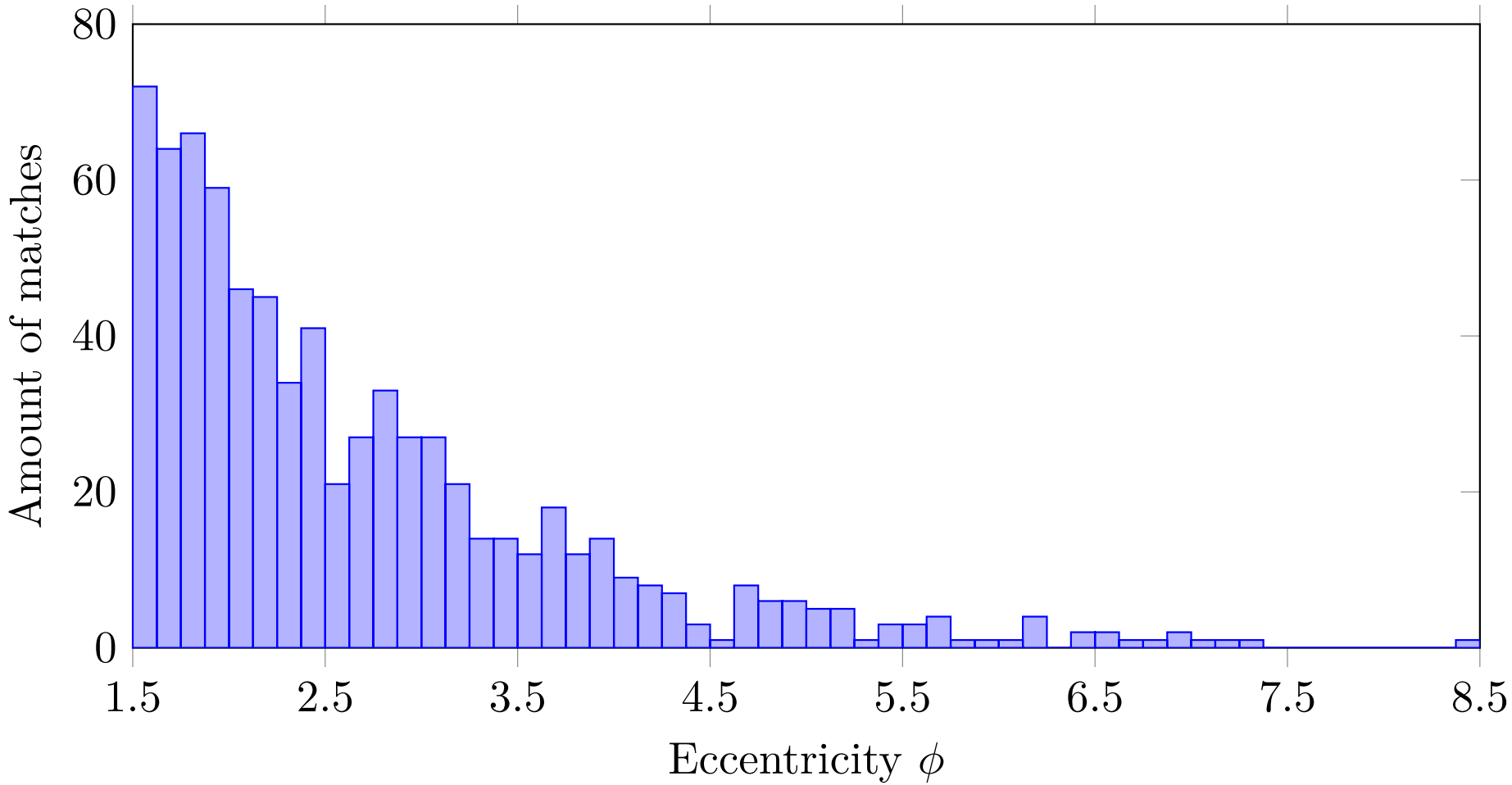
Outline

Approach

Outline

Matches

- 3600 Netflix and 3000 MovieLens users
- a few more than 800 matches, 5 with $\phi > 7$



Validation

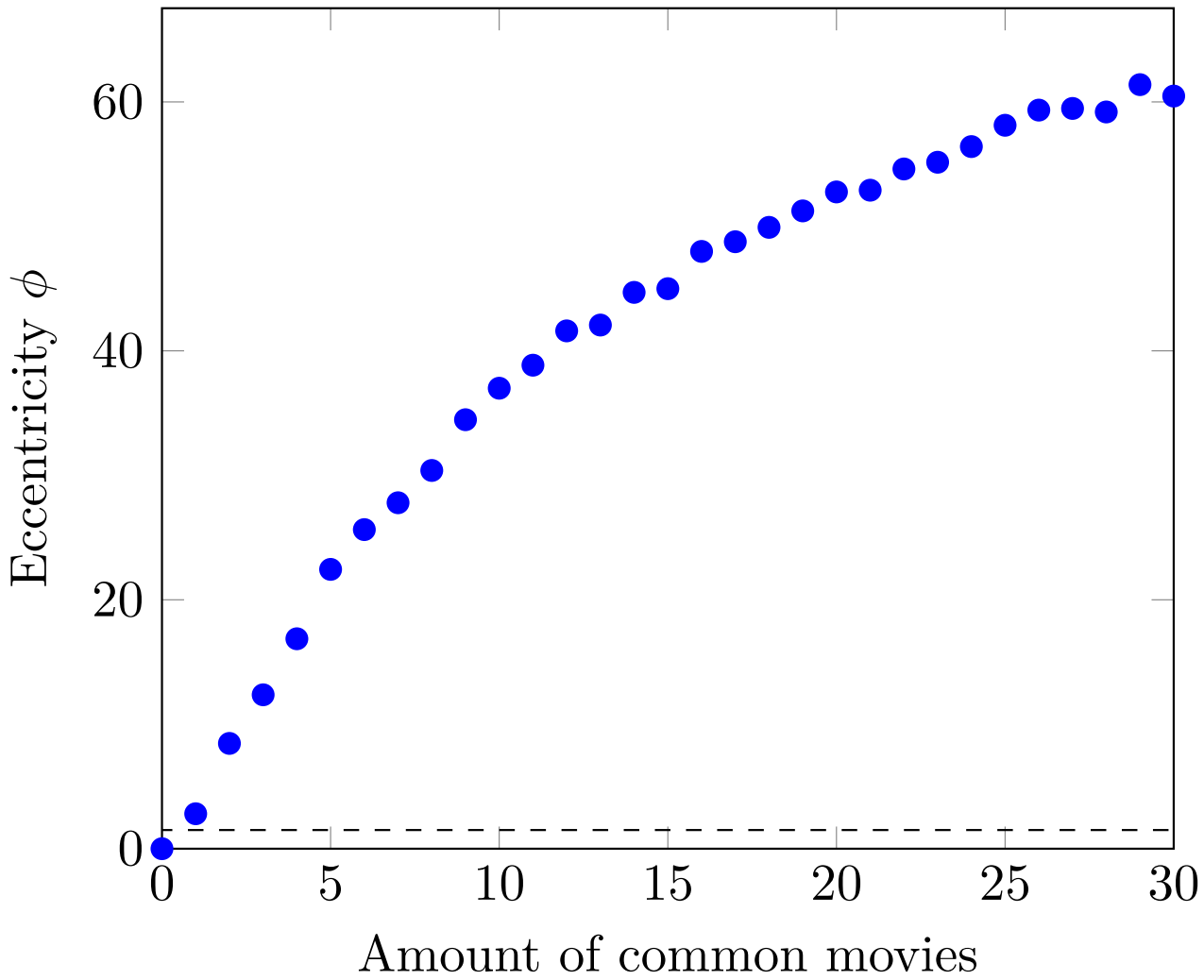
How to validate the results?

- No knowledge of the ground truth of matching users

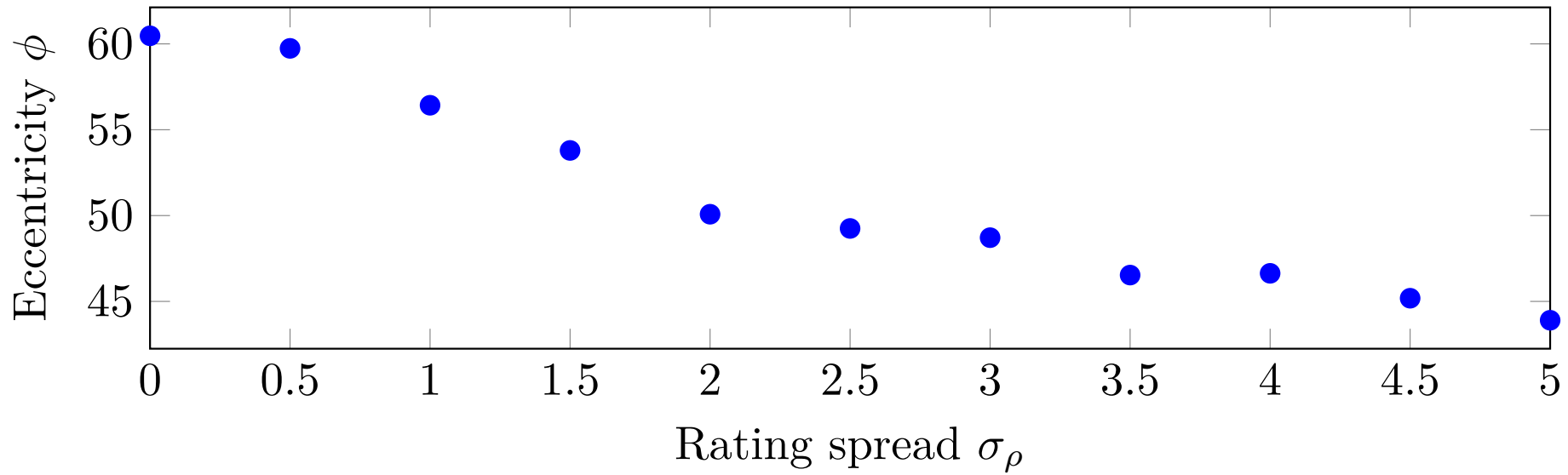
Validation procedure

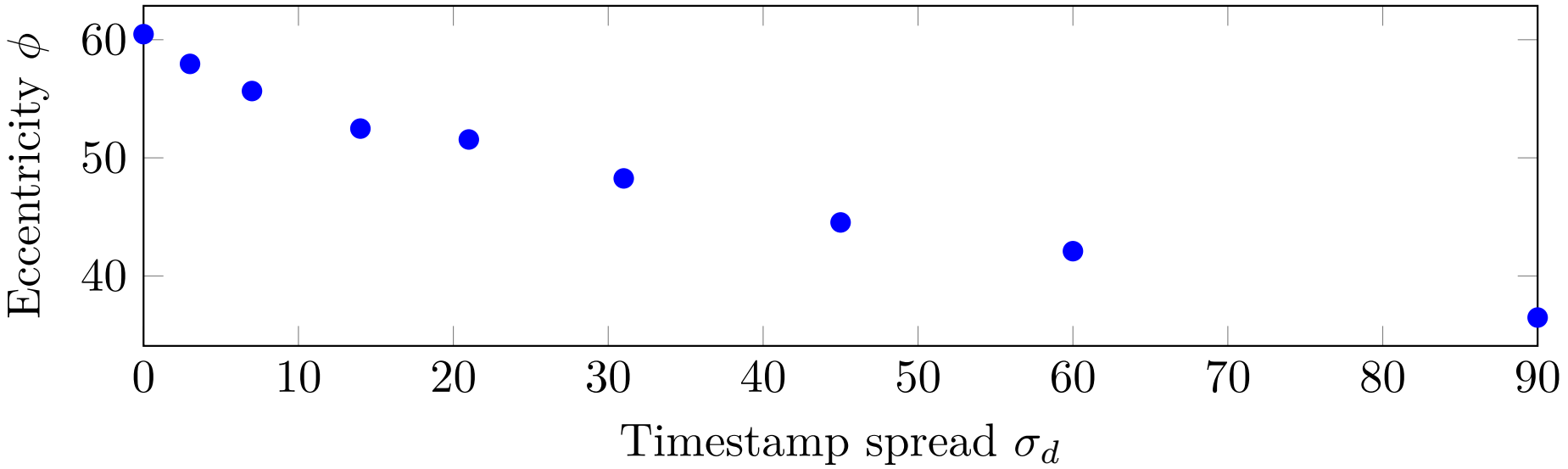
- Put a dummy user in both datasets
- Vary the number of common movies
- Perturb the rating and the timestamp with uniform distributed noise

Validation


Eccentricity ϕ
0
20
40
60
0
5
10
15
20
25
30
Amount of common movies

Robustness


Eccentricity ϕ
60
55
50
45
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Rating spread σ_ρ


Eccentricity ϕ
60
50
40
0
10
20
30
40
50
60
70
80
90
Timestamp spread σ_d

Conclusion

- $\phi_{\max} \approx 8.5$ while the validation process showed $\phi > 30$
- The original Netflix attack: $\phi_{\text{match}} = \{18, 25\}$

$\Rightarrow$ cannot conclude statistical quasi-certainty of de-anonymization

Possible improvements

- Use more than 0.04 % of the possible user combinations
- Tuning of the parameters ϕ, d_0, ρ_0
- Add more features (e.g. movie genres)
- Increase the errors impact on the scoring