Can You Spot the Fakes?
On the Limitations of User Feedback in Online Social Networks

David Mandell Freeman

Head of Anti-Abuse and Anomaly Detection Relevance

LinkedIn

WWW 2017 Perth, Australia

6 April 2017

Fake accounts in social networks

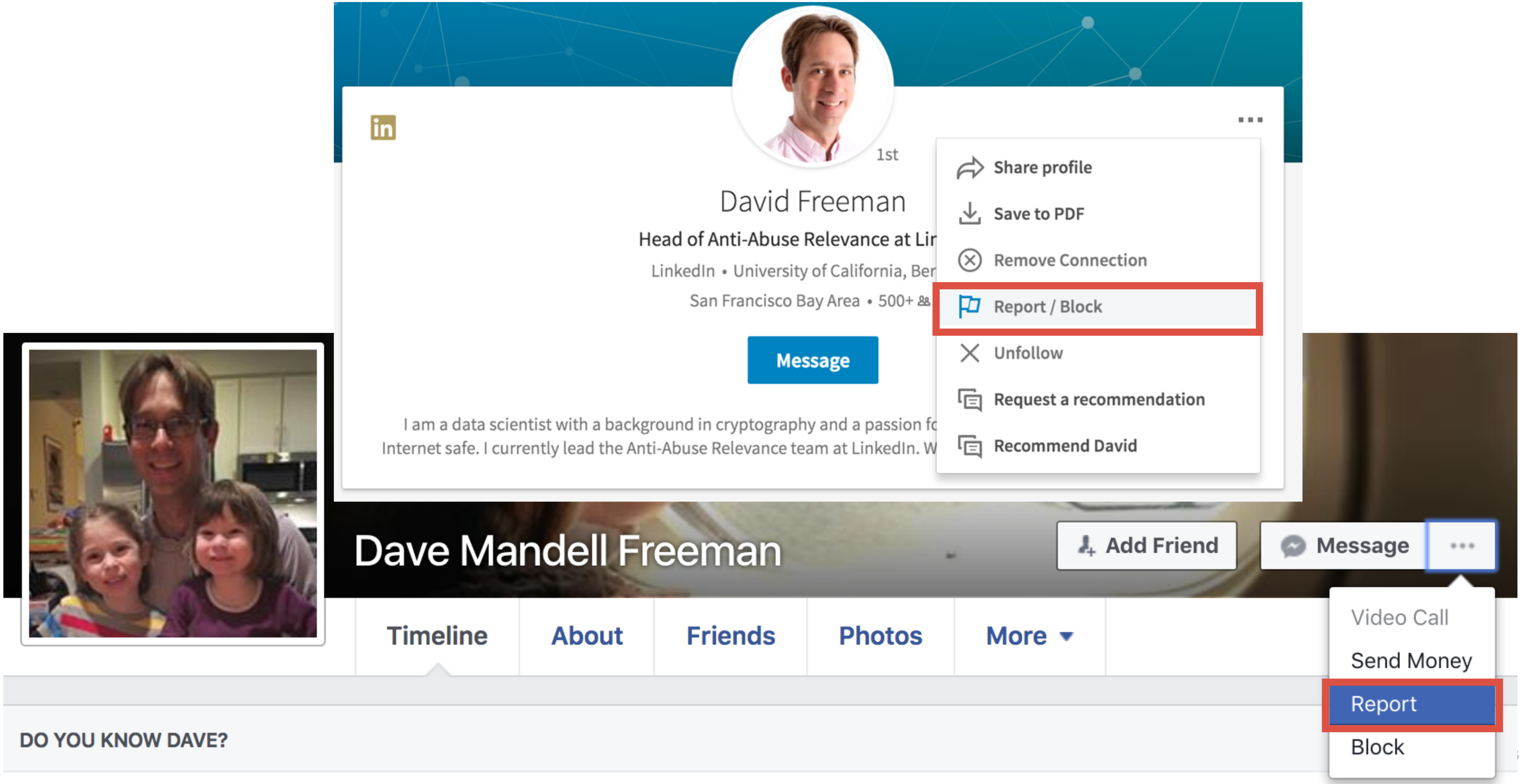
Popular social networks attract bad actors

- scams
- malware
- phishing
- etc.

To carry out abuse, bad guys need fake (or compromised) accounts.

How do we find them?

Reporting fake accounts



Acting on flagging signals

Flagging is a low-precision signal.

- 35% precision in our LinkedIn data set.

Need to accrue multiple flags before taking action.

- This takes time.

We could act faster & more accurately if we knew that some flags were more precise than others.

Research question: is there such a thing a “super-flagger”?

How do we test whether “super-flaggers” exist?

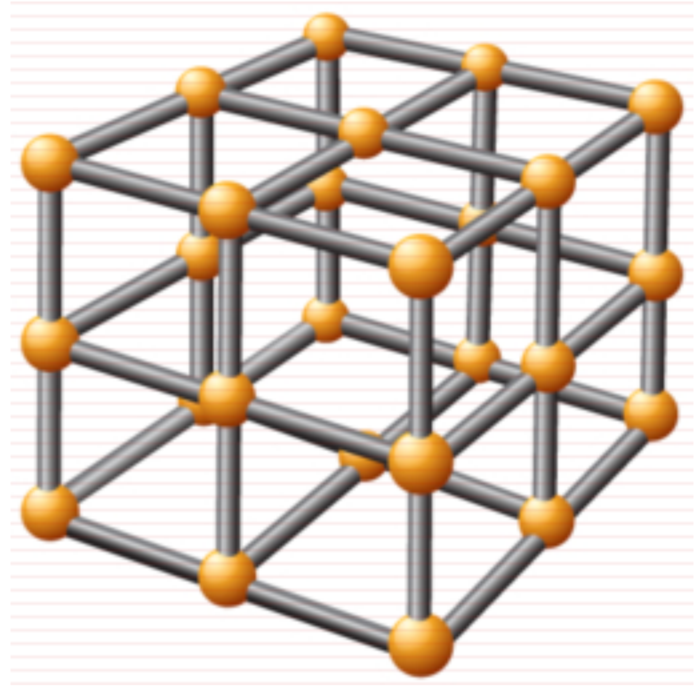
If flagging is a real skill, it must be:

measurable — possible to distinguish from random guessing

repeatable — persistent over repeated sampling

Our contribution

Framework for assessing flagging skill.



Apply framework to LinkedIn data:

- profile report spam
- invitation reject
- invitation accept (signal for *real* accounts)

Conclusion: skilled flaggers exist but are very rare.

- no noticeable impact on metrics

Prior work

[Zheleva et al. ‘08], [Chen et al. ‘15]: Framework to upweight high-precision reporters in spam classification algorithms, mechanism for reputation to evolve.

- Assumes an initial set of high-precision reporters can be identified.
- Assumes identified reporters will continue to be high-precision.

[Wang et al. ’13], [Cresci et al. ’17]: Crowdsourcing studies.

- “People can identify differences between [fake] and legitimate profiles, but most individual testers are not accurate enough to be reliable.”
- Low accuracy on “social spambots”

[Moore-Clayton ‘08] [Chia-Knapskog ’11]: “wisdom of crowds”

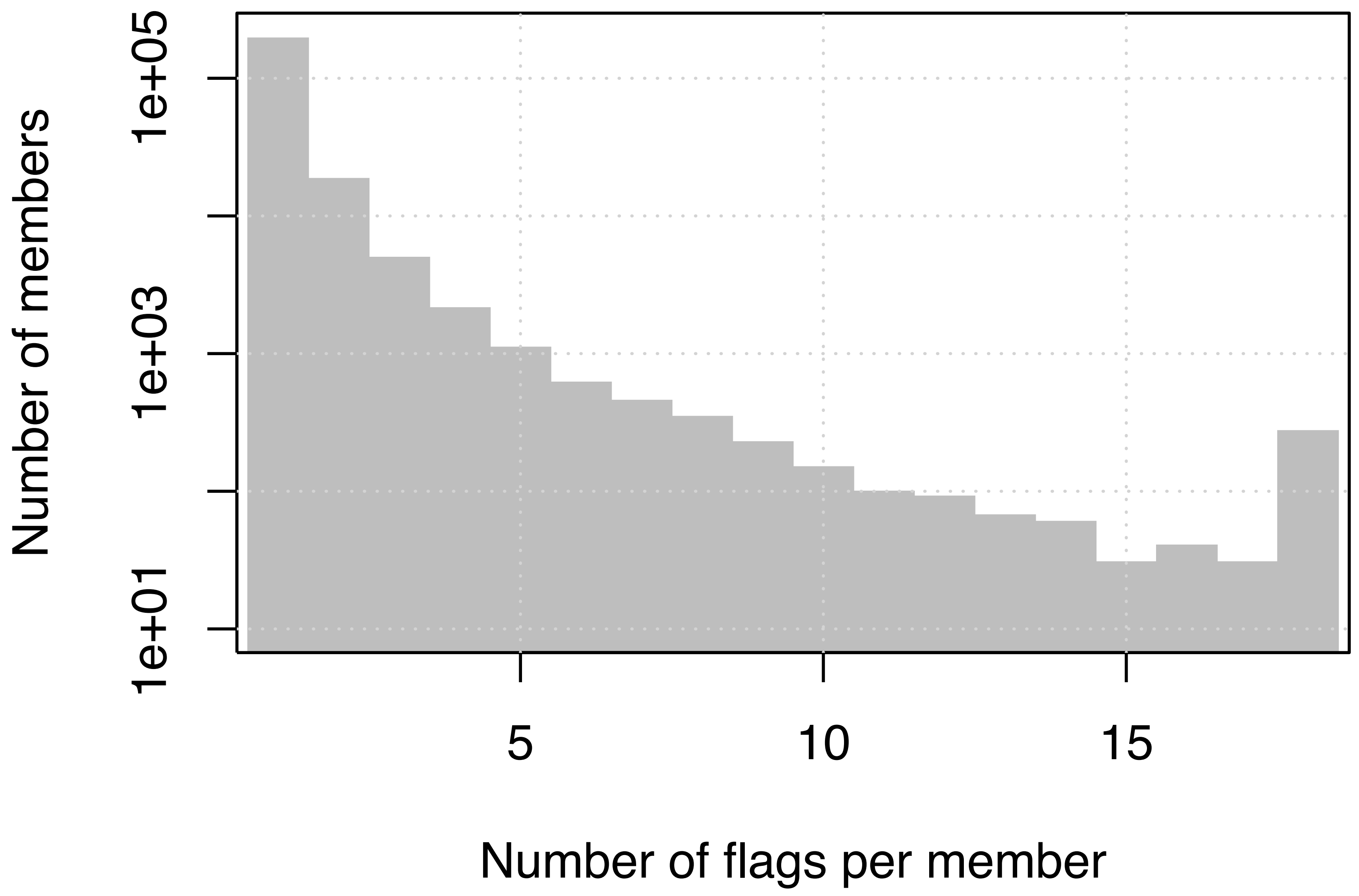
- Frequent reporters have higher accuracy (counter to our findings)

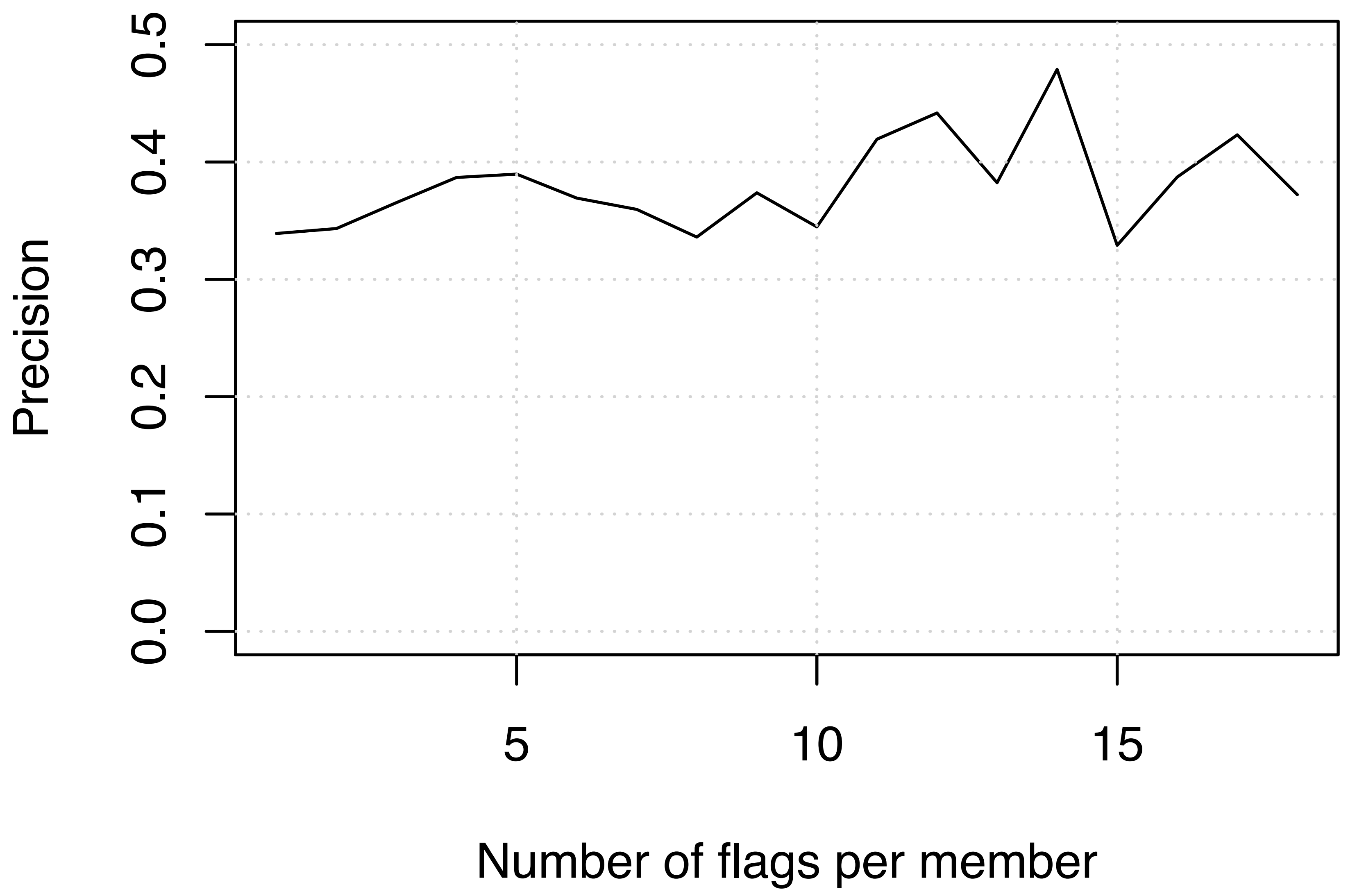
Profile flagging data set

Data: all LinkedIn “fake profile” flags over 6-month period

- 293K flags, 227K reporters, 238K reports
- Anti-Abuse team labeled flagged accounts as real or fake
- 35% overall precision

Precision does not improve with number of flags:





(last bucket is all members with ≥18 flags)

Measurability: Precision

How many flags did the user get right?

$$P(u) = \frac{\#\text{ correct flags}}{\#\text{ flags}}$$

Problem: insensitive to number of flags

- 1 out of 1 is as good as 50 out of 50

Solution: smoothing

$$P_s(u) = \frac{\#\text{ correct flags} + \alpha}{\#\text{ flags} + 2\alpha}$$

- find α by optimizing on a test set

Measurability: Precision

How many flags did the user get right?

$$P(u) = \frac{\text{\# correct flags}}{\text{\# flags}}$$

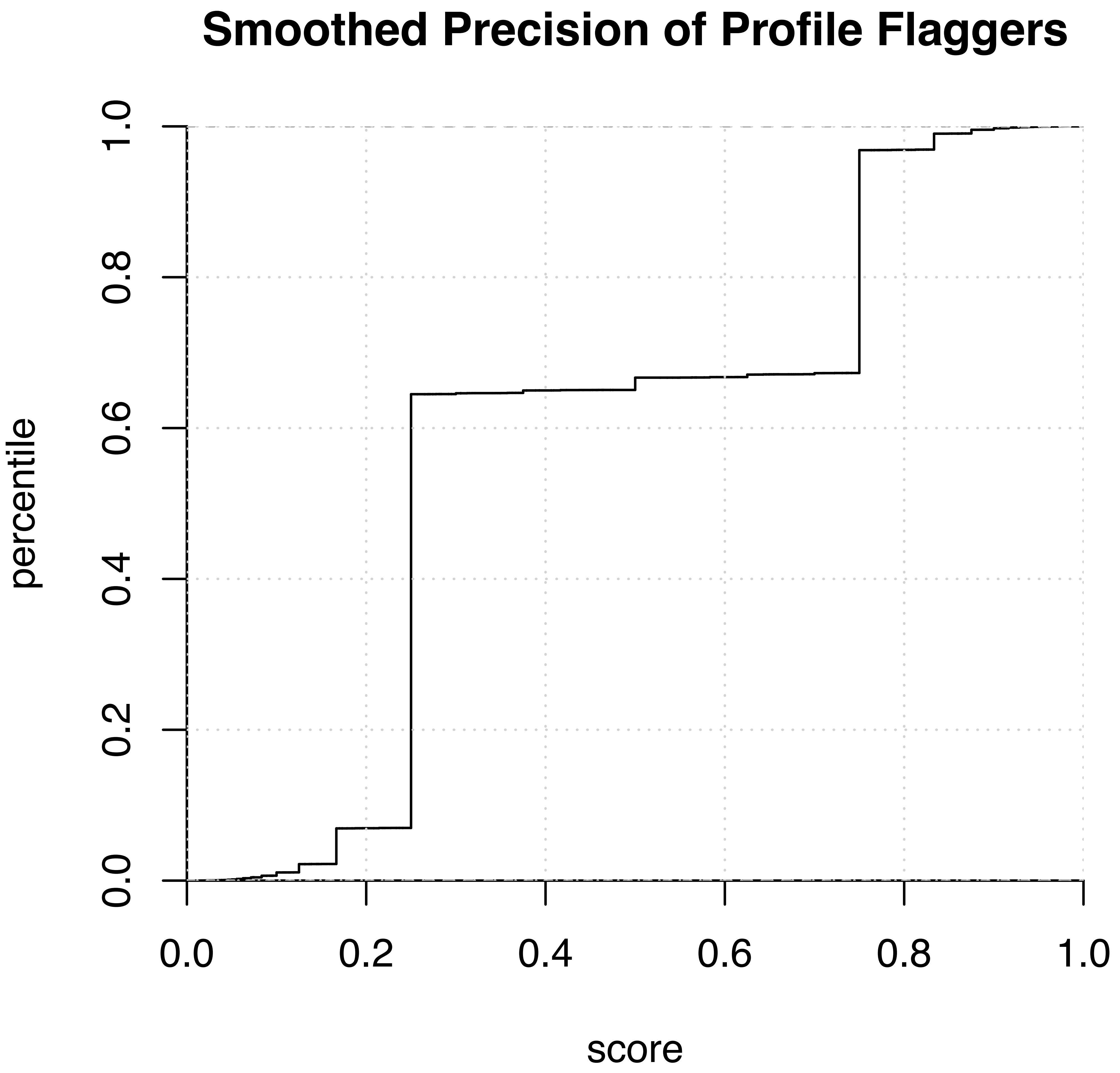
Problem: insensitive to number of flags

- 1 out of 1 is as good as 50 out of 50

Solution: smoothing

$$P_s(u) = \frac{\text{\# correct flags} + \alpha}{\text{\# flags} + 2\alpha}$$

- find α by optimizing on a test set



Measurability: Informedness

u	Report	Ignore
Real	5	5
Fake	5	5

$$\text{precision} = \frac{5}{10} = 0.5$$

u'	Report	Ignore
Real	5	95
Fake	5	5

$$\text{precision} = \frac{5}{10} = 0.5$$

Measurability: Informedness

Precision is insensitive to level of fake account exposure:

u	Report	Ignore
Real	5	5
Fake	5	5

$$\text{precision} = \frac{5}{10} = 0.5$$

u'	Report	Ignore
Real	5	95
Fake	5	5

$$\text{precision} = \frac{5}{10} = 0.5$$

Measurability: Informedness

Precision is insensitive to level of fake account exposure:

u	Report	Ignore
Real	5	5
Fake	5	5

$$\text{precision} = \frac{5}{10} = 0.5$$

u'	Report	Ignore
Real	5	95
Fake	5	5

$$\text{precision} = \frac{5}{10} = 0.5$$

Informedness: How much better is the user at flagging fake accounts than real ones?

$$I(u) = \text{TPR} - \text{FPR} = \frac{\#\text{ flags of fakes}}{\#\text{ fakes seen}} - \frac{\#\text{ flags of reals}}{\#\text{ reals seen}}$$

Measurability: Informedness

Precision is insensitive to level of fake account exposure:

u	Report	Ignore
Real	5	5
Fake	5	5

u'	Report	Ignore
Real	5	95
Fake	5	5

$$\text{precision} = \frac{5}{10} = 0.5 \qquad \text{precision} = \frac{5}{10} = 0.5$$

$$\text{informedness} = \frac{5}{10} - \frac{5}{10} = 0 \qquad \text{informedness} = \frac{5}{10} - \frac{5}{100} = 0.45$$

Informedness: How much better is the user at flagging fake accounts than real ones?

$$I(u) = \text{TPR} - \text{FPR} = \frac{\#\text{ flags of fakes}}{\#\text{ fakes seen}} - \frac{\#\text{ flags of reals}}{\#\text{ reals seen}}$$

Measurability: Informedness

Precision is insensitive to level of fake account exposure:

u	Report	Ignore
Real	5	5
Fake	5	5

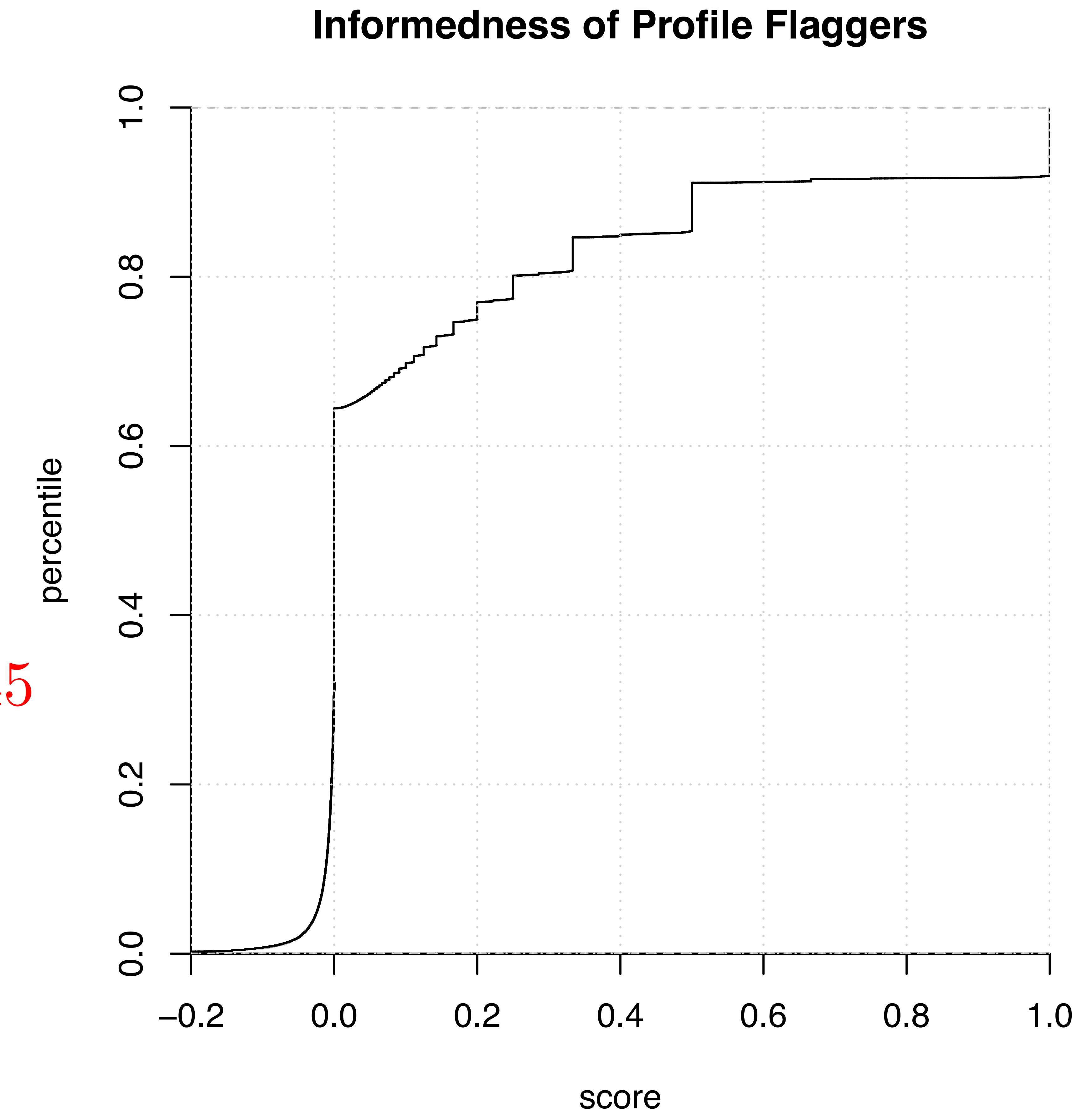
u'	Report	Ignore
Real	5	95
Fake	5	5

$$\text{precision} = \frac{5}{10} = 0.5 \qquad \text{precision} = \frac{5}{10} = 0.5$$

$$\text{informedness} = \frac{5}{10} - \frac{5}{10} = 0 \qquad \text{informedness} = \frac{5}{10} - \frac{5}{100} = 0.45$$

Informedness: How much better is the user at flagging fake accounts than real ones?

$$I(u) = \text{TPR} - \text{FPR} = \frac{\#\text{ flags of fakes}}{\#\text{ fakes seen}} - \frac{\#\text{ flags of reals}}{\#\text{ reals seen}}$$



Is it skill or luck?

v	Report	Ignore
Real	2	2
Fake	1	0

v'	Report	Ignore
Real	20	20
Fake	10	0

$$\text{informedness} = \frac{1}{1} - \frac{2}{4} = 0.5 \qquad \text{informedness} = \frac{10}{10} - \frac{20}{40} = 0.5$$

Use a statistical hypothesis test to distinguish the two!

Fisher's exact test on the 2 x 2 contingency table.

Null hypothesis: user is equally likely to flag real and fake accounts.

p-value: probability of finding a matrix "at least as extreme" as *M.*

Is it skill or luck?

v	Report	Ignore
Real	2	2
Fake	1	0

v'	Report	Ignore
Real	20	20
Fake	10	0

$$\text{informedness} = \frac{1}{1} - \frac{2}{4} = 0.5 \qquad \text{informedness} = \frac{10}{10} - \frac{20}{40} = 0.5$$

$$p = 1 \qquad p = 0.003$$

Use a statistical hypothesis test to distinguish the two!

Fisher's exact test on the 2 x 2 contingency table.

Null hypothesis: user is equally likely to flag real and fake accounts.

p-value: probability of finding a matrix "at least as extreme" as *M.*

Measurability: Hypothesis Testing

Fisher's test produces a *p*-value: probability of finding a matrix "at least as extreme" as *M.*

— define "Fisher Score" = 1 – *p*-value

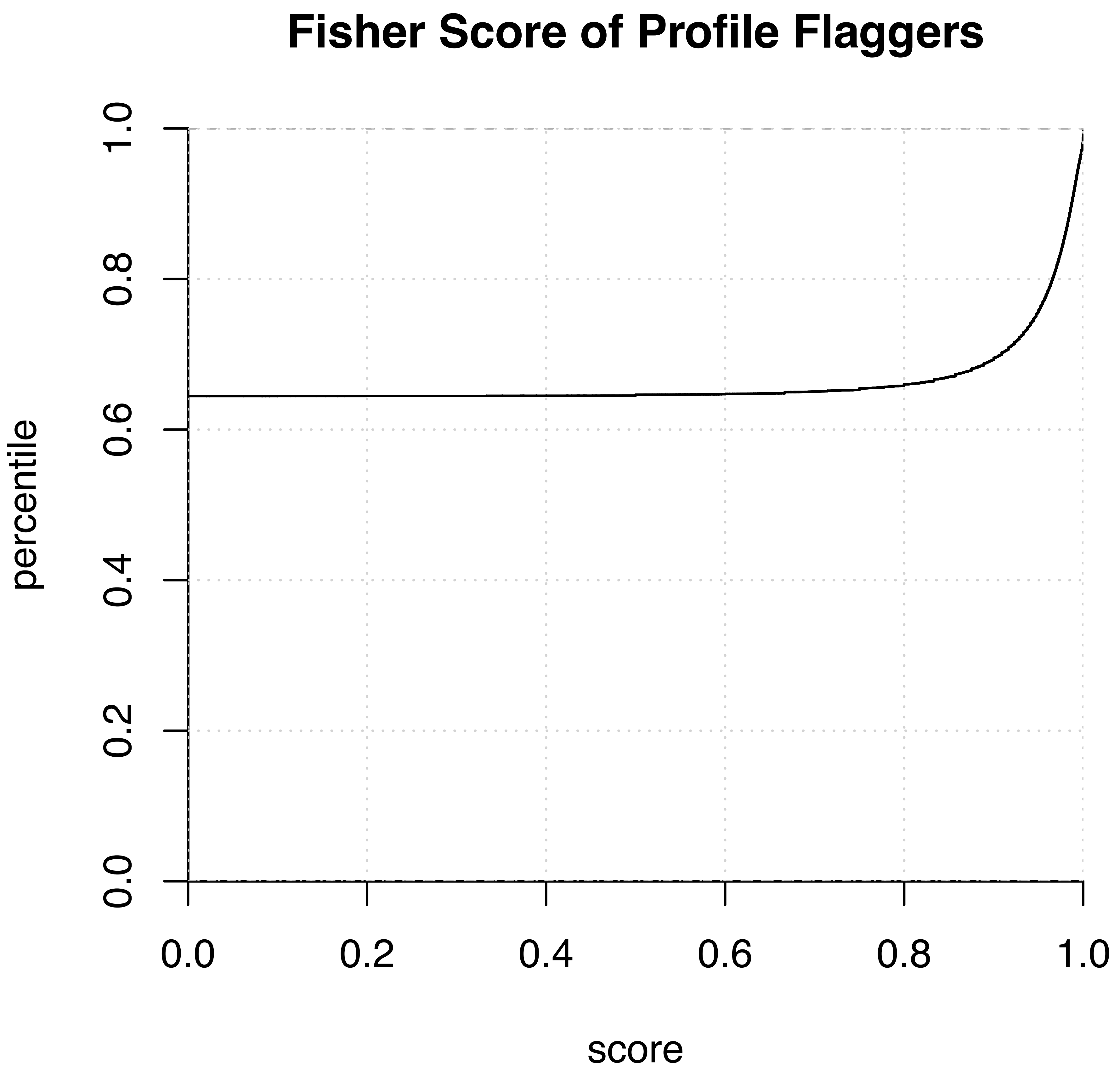
Problem: statistically significant flaggers may not be good flaggers

w	Report	Ignore
Real	20	80
Fake	5	5

$$\text{precision} = \frac{5}{25} = 0.2$$

$$\text{informedness} = \frac{5}{5} - \frac{20}{100} = 0.3$$

$$\text{Fisher score} = 1 - 0.05 = 0.95$$



Repeatability — Correlation

Are skilled flaggers in data set *A* the same as skilled flaggers in data set *B*?

Pearson correlation coefficient: linear correlation of scores.

Spearman correlation coefficient: Pearson correlation of rank vectors.

Flagging Score	Pearson	Spearman
Smoothed Precision	0.69	0.66
Informedness	0.52	0.49
Fisher Score	0.62	0.63

Problem: independent of score magnitude

user	A score	B score
a	0.94	0.1
b	0.95	0.2
c	0.96	0.3
d	0.97	0.4
e	0.98	0.5

Perfect correlation!

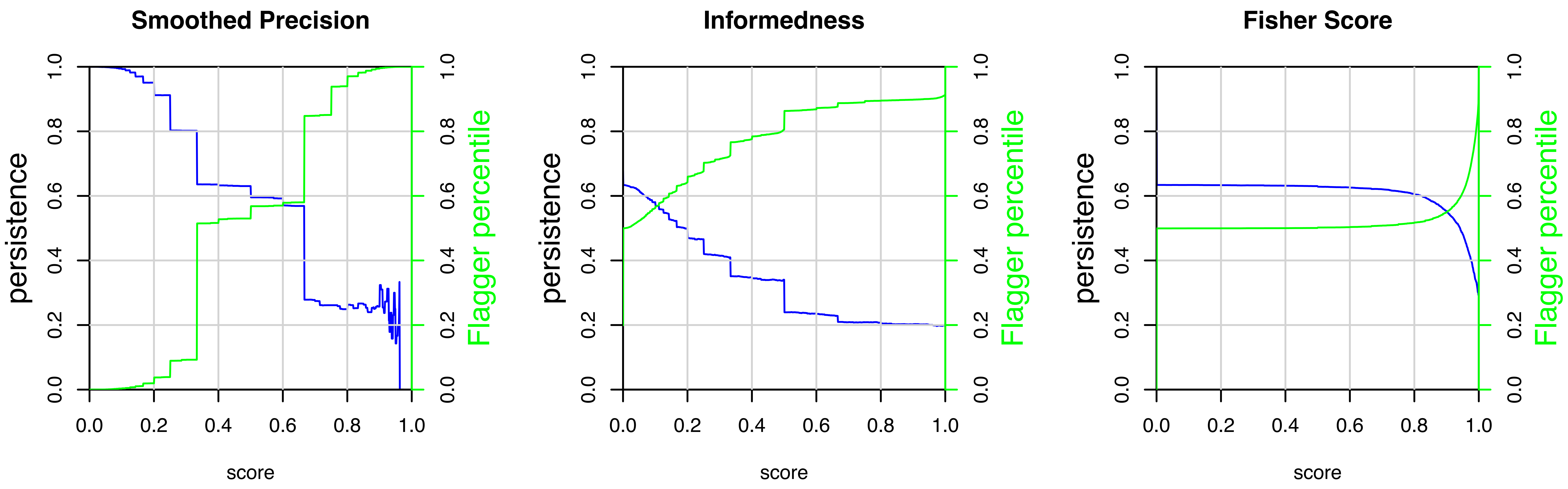
Repeatability — Persistence

Probability that user with a good score in data set *A* also has a good score in data set *B*?

Define *persistence at score* β to be

$$\pi(\beta) = \frac{\#\text{ users with score} > \beta \text{ in } A \textit{ and } B}{\#\text{ users with score} > \beta \text{ in } A \textit{ or } B}$$

Persistence on flagging data:



Putting it all together

Compute skill threshold for each measurement based on precision on a held-out test set.

- Threshold is such that error rate is less than half the average.

Define "skilled flagger" to be one who is above the threshold on **2 of 3 metrics**, on **2 different data sets**

- high smoothed flagging precision
- flags real and fake accounts in different proportion
- difference in behavior in flagging real and fake accounts is statistically significant

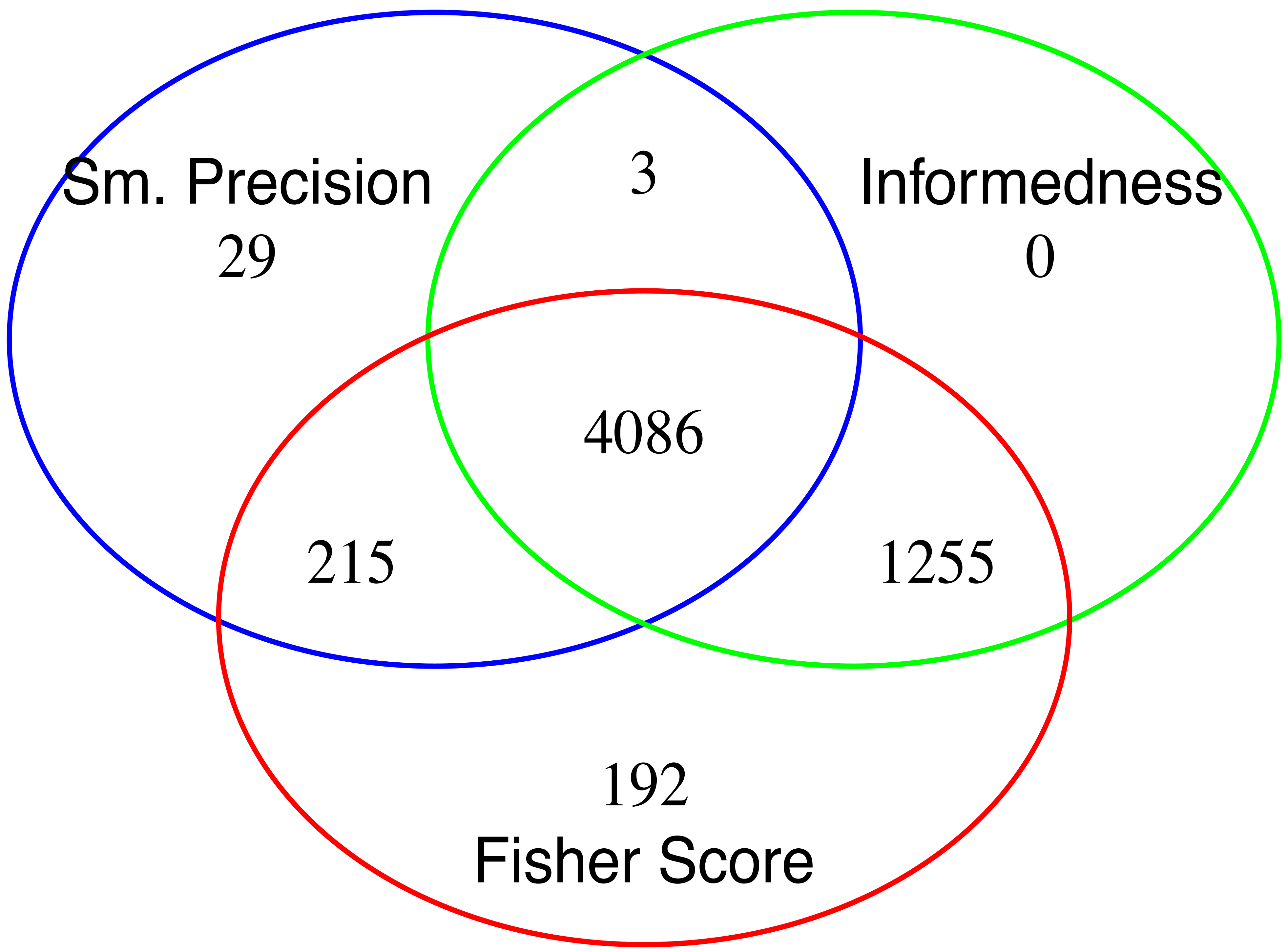
Profile flagging — skilled flaggers

5600 skilled flaggers

- 31% of those who flagged ≥2 times
- 2.4% of all flaggers
- 82% cumulative precision

4300 high-precision skilled flaggers

- 13940 accounts flagged **(77/day)**
- 97% cumulative precision



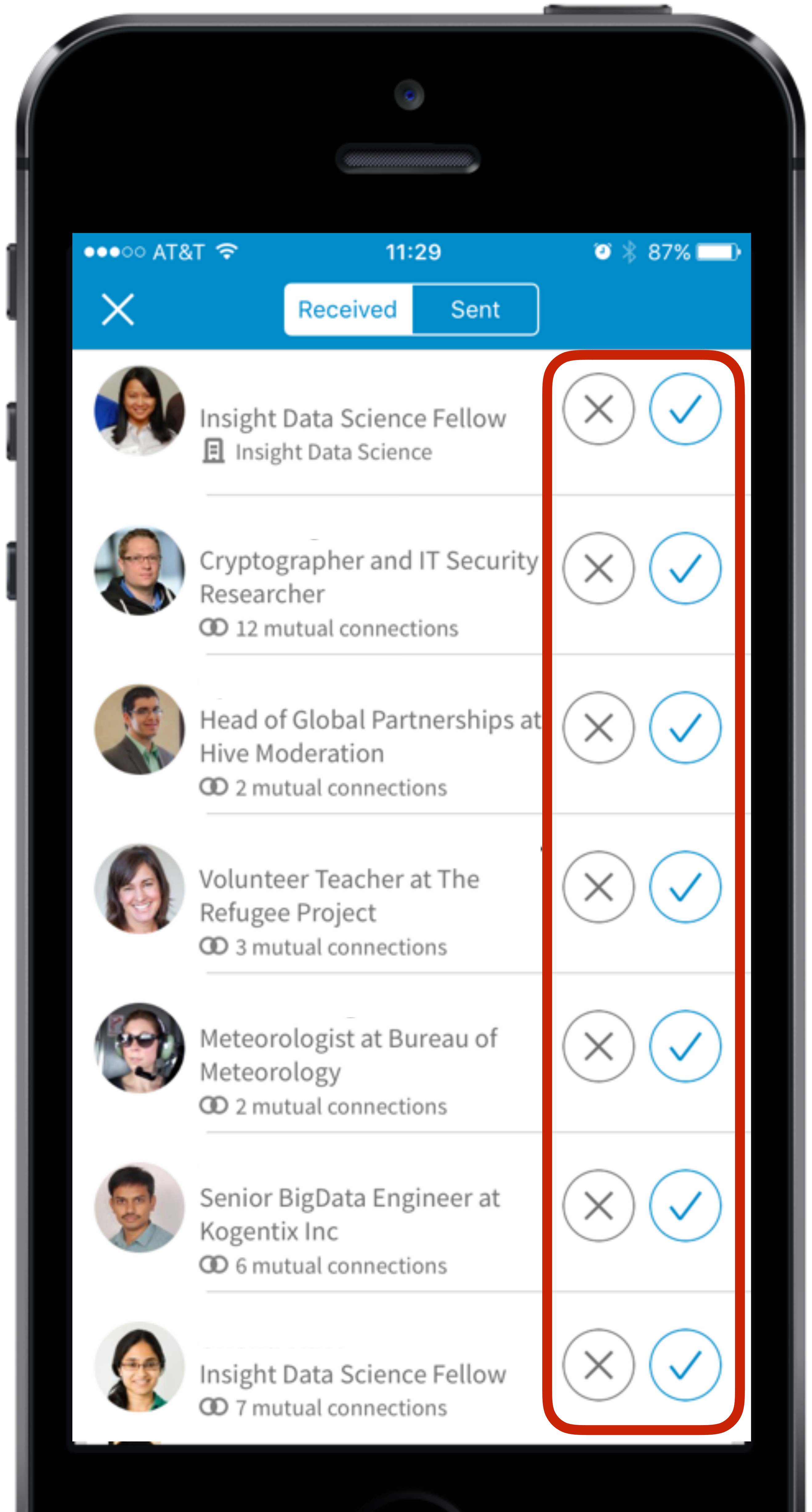
Data set 2: Invitation response

Invitation *reject*: reporting signal on *fake* accounts

Invitation *accept*: reporting signal on *real* accounts

Evaluation:

- 500,000 members from June 2016 receiving ≥2 spam and ≥3 non-spam invitations
- look at responses within the first 24 hours
- 1.3% were skilled at *rejecting fakes*
- 3.8% were skilled at *accepting reals*



An experiment

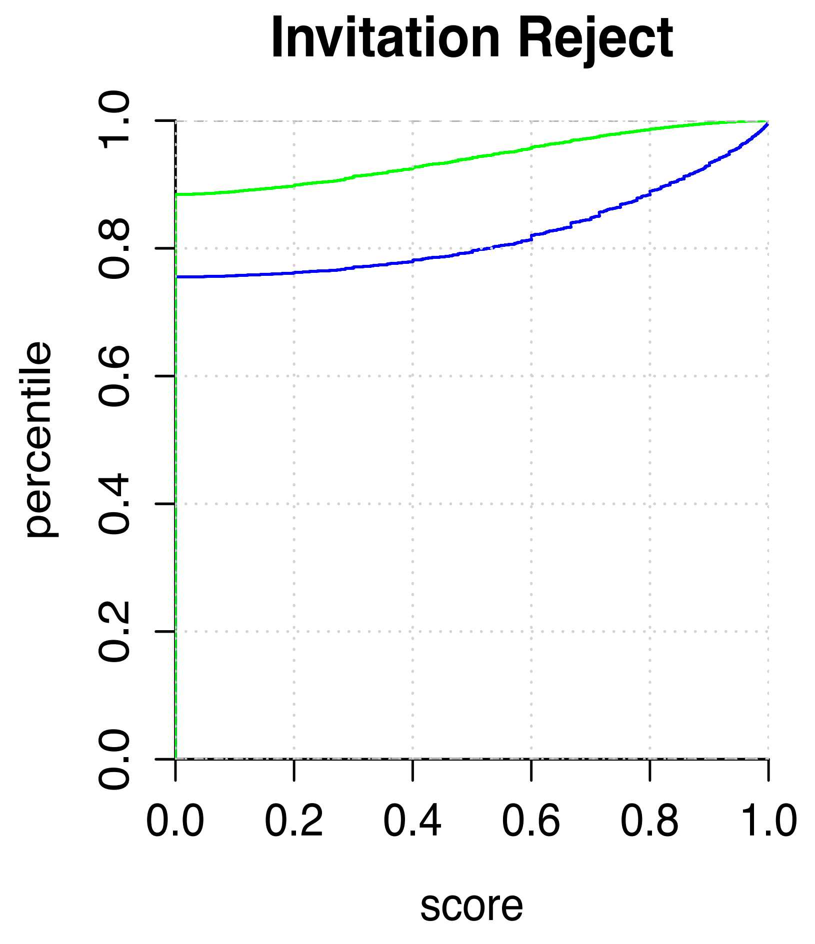
Simulation: replace member's responses to *fake* accounts with binomial samples distributed like responses to *real* accounts.

	Report	Ignore
Real	5	20
Fake	8	2
Simulated Fake $\sim B(0.25)$	2	8

$p = 0.002$

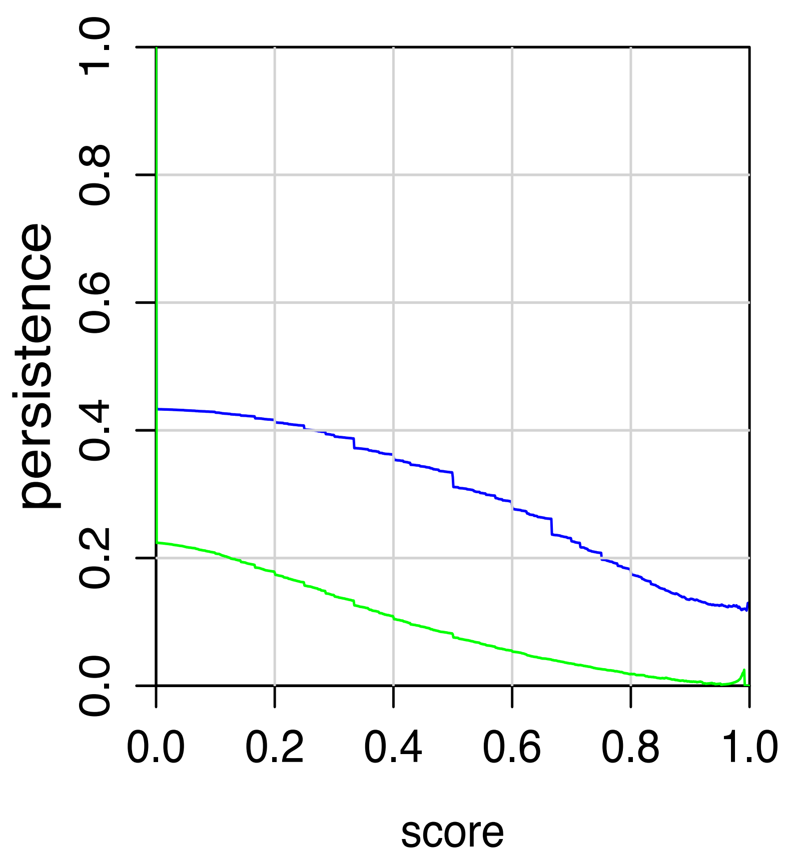
$p = 1$

- Fisher scores are lower for simulated data
- persistence drops to zero much more quickly for simulated data



Fisher score distribution

Blue = real
Green = simulated



Fisher score persistence

Conclusions

Motivating question: *Are there some social network users who are good at identifying fake accounts?*

Answer: yes, but not enough to make acting on the signal worthwhile:

- < 2.4% of profile flaggers
- < 1.3% of members rejecting invitations
- < 3.8% of members accepting invitations (i.e. identifying real accounts)

Further work:

- investigate UI changes to improve flagging ability
- find other features correlated with skill (e.g. geo)

Questions?

dfreeman@linkedin.com