Searching For Spam: Detecting Fraudulent Accounts Via Web-Search

Marcel Flores and Aleksandar Kuzmanovic
Northwestern University

Twitter Spam

- Twitter presents fresh challenges:
  - Forced brevity,
  - easily obscured content,
  - and non-symmetric social links.

Example

Joel Nickel @joelnickel Mar 10

Jon Stewart Trashes CNN Again & Again on 'Larry King Live';
youtu.be/K_qJiReI8hU

Details

stratfordun7
@stratfordun7

Follow

@joelnickel nugangxin.info/DzKmrI

Reply Retweet Favorite More

2:02 PM - Mar 10, 2013

Existing Techniques

- Generally consider:
  - Message format
  - Message content
  - Social Graph Location

Require time!

Our Approach

- Users often use many interlinking sites
  - OSNs, blogs, forums
  - Often use similar names
- Spam accounts are often throw-aways

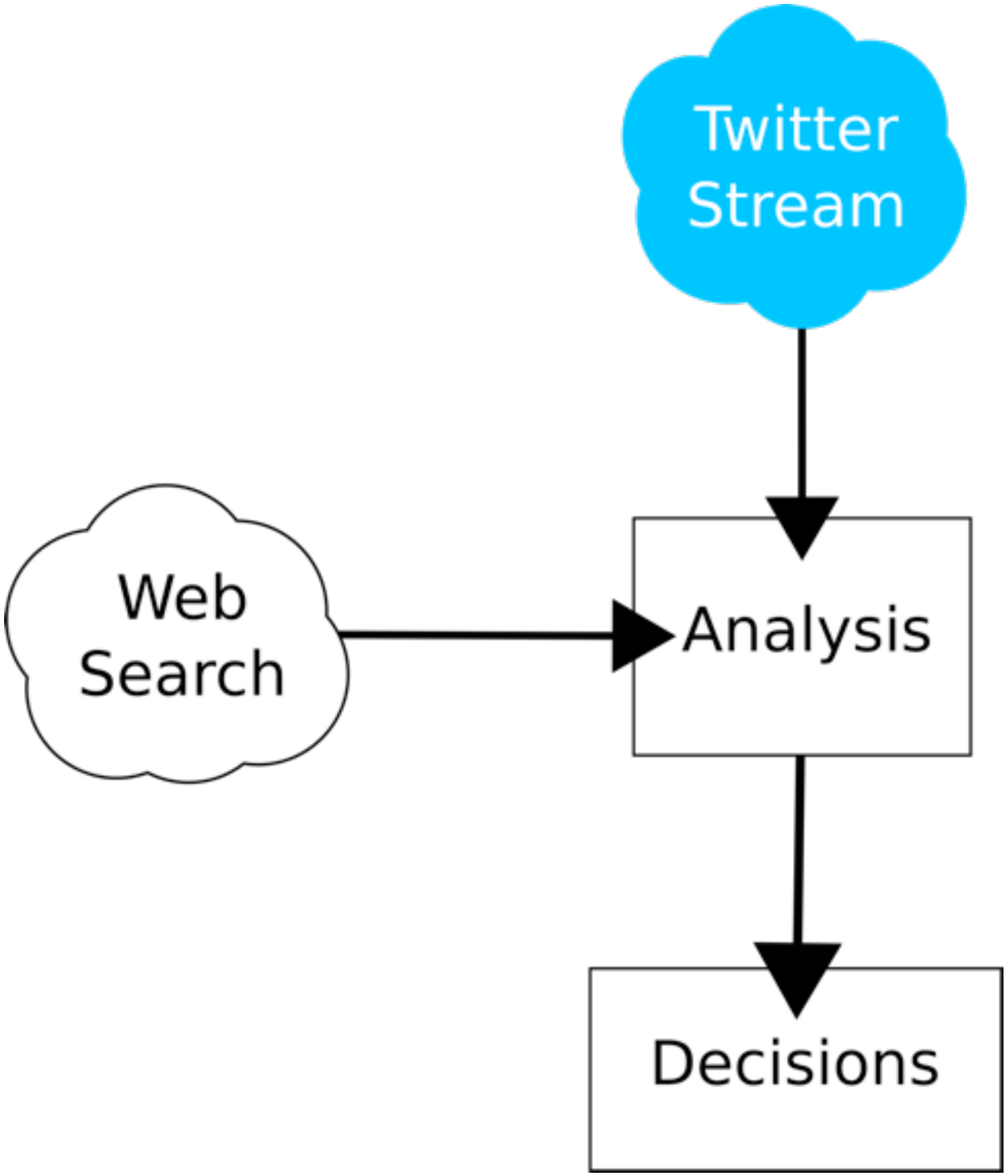
Our Approach

We can measure this distributed online presence with a web search!

Our Approach

- Can be done with existing indices.
- Mimicking the effect would be very difficult.
- Very fast, account need not have generated any content.
- Could detect fraudulent accounts at creation time.

So how does it all work?


Twitter
Stream
Web
Search
Analysis
Decisions

Methods

- Perform a web search for the username and display name.
- Eliminate noise in the results:
  - Remove Twitter and Twitter Services,
  - Remove frequent results.
- White-list a set of known-helpful sites.

Methods

- If there are results left, declare the account legitimate.

How well does it work?

Dataset

- Collect over 20 GB of data from the “trickle.”
- Filter out non-English.
- Save profile information for every unique account seen which performed an @ mention
  - 110,000 total accounts.
- Perform web searches for each account.

Verification Labeling

- Check account status 2 weeks after:
  - Suspended indicates spam
- 21.25% of observed accounts were suspended.

Verification Labeling

- Perform a manual check of 200 randomly sampled un-suspended accounts:
  - 18% are clearly fraudulent
  - Will inflate our false positive rate

Performance

- We are able to achieve:
  - True positive rate: 74.23%
  - False positive rate: 10.67%

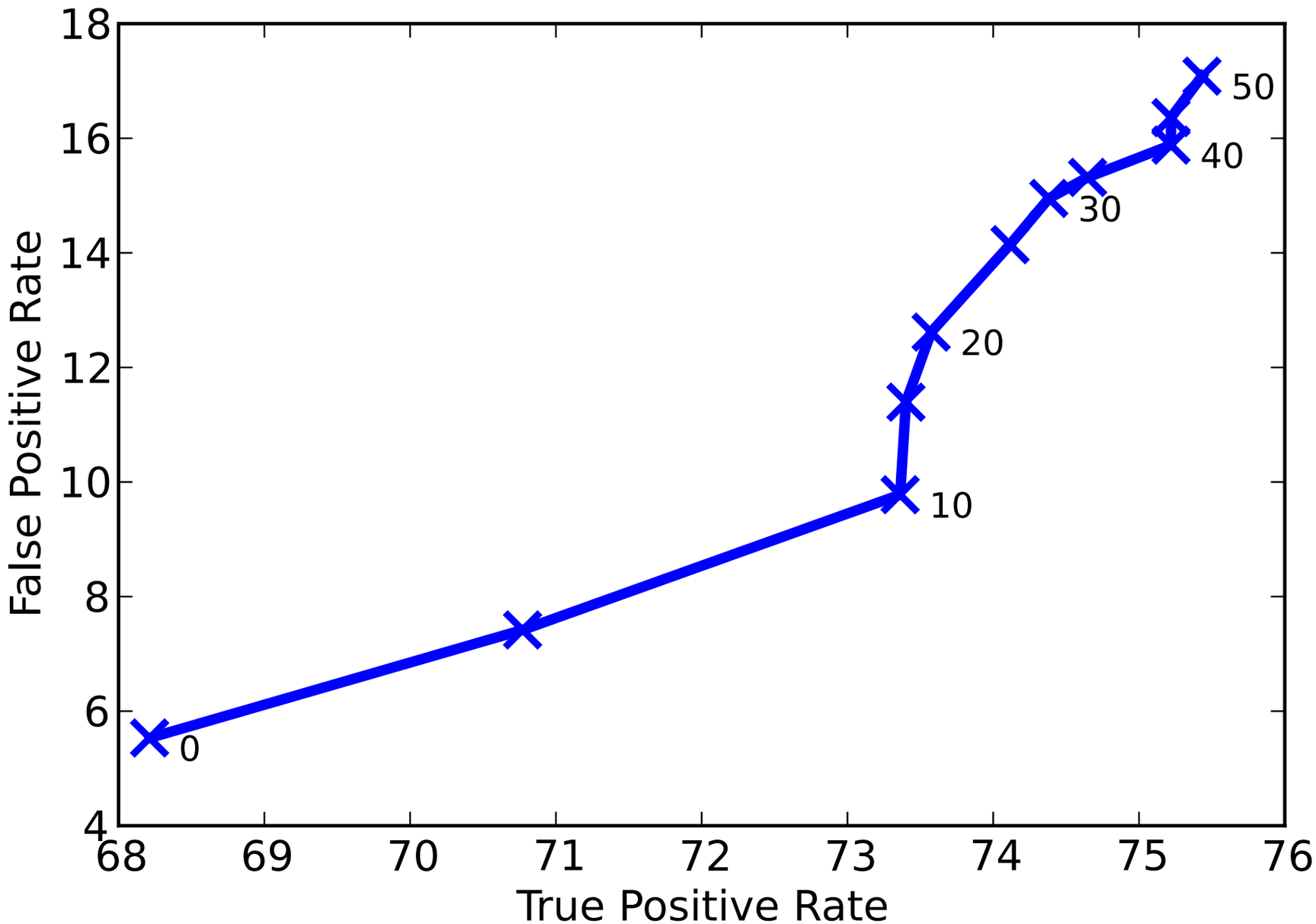
False Positives

- Manually inspect 200 false positives
  - 61% clearly fraudulent
  - 7.5% appeared compromised
- May have:
  - TPR 79.2% FPR 4.5%

Noise Reduction Parameters

- How long should our blacklist of frequent results be?

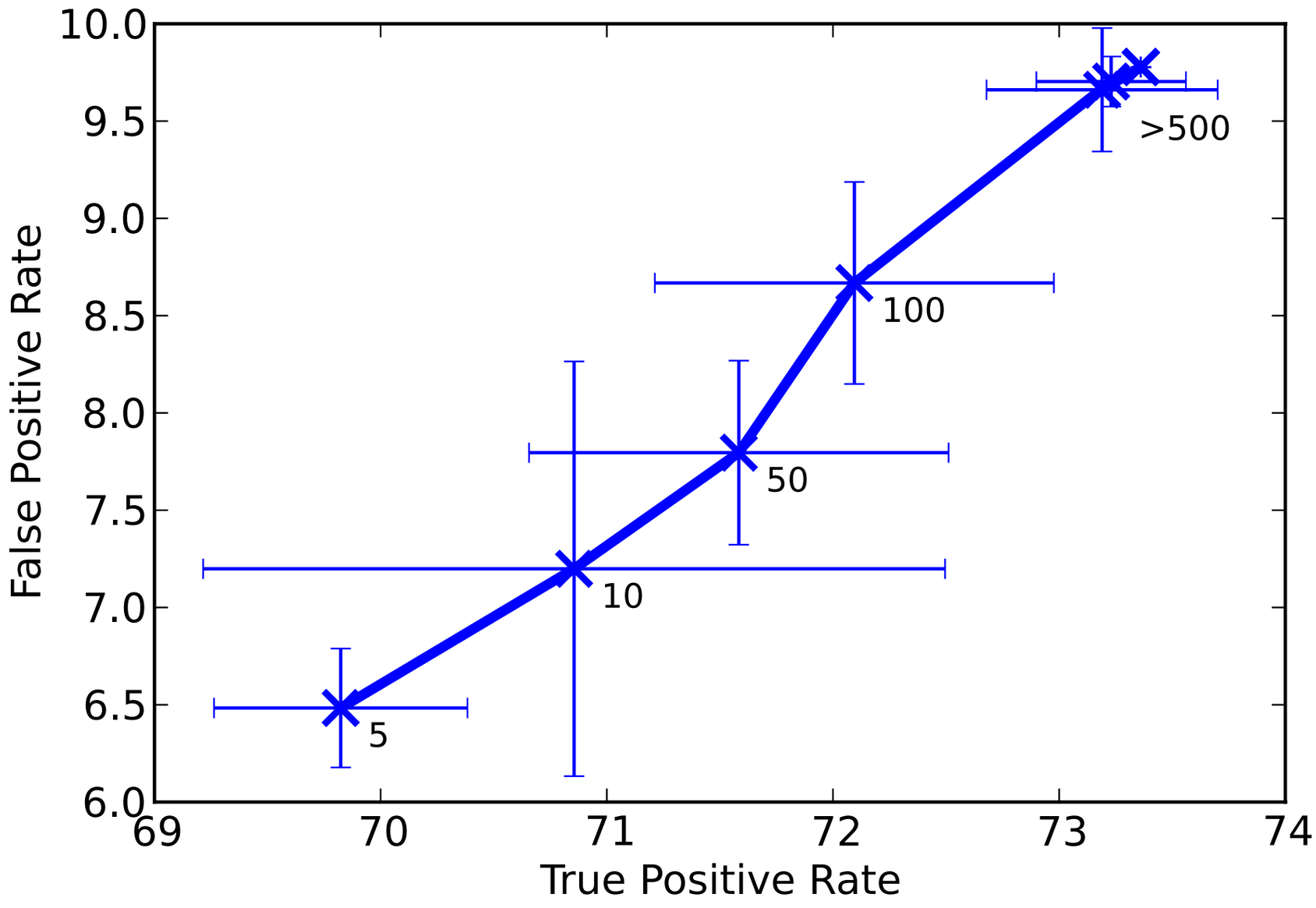
Tuning


False Positive Rate
True Positive Rate
18
16
14
12
10
8
6
4
68
69
70
71
72
73
74
75
76
0
10
20
30
40
50

How long does it take?

- How many search result sets must we see to build an effective list?

Training Speed



Conclusion

- Makes call on the nature of an account using a measure of their web presence.
- Stands to work well as a first step in a comprehensive system.
- Achieve a TPR of 74.67%
- System is straightforward and works quickly.

Conclusion

- Data and tools are available at:
  - http://users.eecs.northwestern.edu/~mef294/projects/twitter.html

Questions?