




How to find spam on Twitter?

Mourjo Sen

Under the guidance of
Arnaud Legout, Maksym Gabielkov



Outline

- ◎ **Background, problem statement, workflow**
 - ◎ Definition of our metric of trust
 - ◎ Spam detection methodology
 - ◎ Testing our method
 - ◎ Conclusion: Next Steps
- 

#JeSuisCharlie

Mentioned 6,500 times per minute
3.4 million times in a day



The dark side of social media

A decorative network diagram in the top right corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue and others in grey.

The dark side of social media

- ◎ A hacker starts an online rumour about a plane crash on Twitter

The dark side of social media

- ◎ A hacker starts an online rumour about a plane crash on Twitter
- ◎ The “news” goes viral and the airline’s stock plummets

The dark side of social media

- ◎ A hacker starts an online rumour about a plane crash on Twitter
- ◎ The “news” goes viral and the airline’s stock plummets
- ◎ The hacker makes a fortune on stock short sales

**Eleazar David Melendez**

Become a fan



Eleazar.Melendez@huffingtonpost.com

Twitter Stock Market Hoax Draws Attention Of Regulators

Posted: 02/01/2013 6:38 pm EST | Updated: 02/03/2013 11:35 am EST

From the boiler room-basement brokerages of southern New Jersey to the opulent office suites of midtown Manhattan hedge funds, the U.S. financial police are supposed to track down stock market fraud wherever it takes place. For what appears to be the first time, a tip-off is leading them into the world of Twitter.

U.S. market regulators are trying to determine if a message posted on a hoax Twitter account this week was used as part of a securities fraud scheme, The Huffington Post has learned. The inquiry will be looking at tweets, sent Tuesday from an account thinly disguised as that of a well-known equity research group, that contained false information regarding Silicon Valley company Audience Inc. The tweets caused a violent sell-off in the stock of that company Tuesday afternoon, which some market participants noted was likely intensified by high-frequency trading robots.

Diane Vanasse, a spokesperson for Audience, Inc. said the "company was certainly aware of the hoax tweets" and said that "Nasdaq is investigating" the matter. Joe Christinat, a spokesperson for exchange operator Nasdaq OMX, said the electronic trading board "did refer the matter to FINRA."

Real-world influence of Twitter



Real-world influence of Twitter

◎ Political campaigns





Real-world influence of Twitter

- ◎ Political campaigns
- ◎ Marketing campaigns + promotions





Real-world influence of Twitter

- ◎ Political campaigns
 - ◎ Marketing campaigns + promotions
 - ◎ Stock markets
- 




Real-world influence of Twitter

- ◎ Political campaigns
- ◎ Marketing campaigns + promotions
- ◎ Stock markets
- ◎ Journalism: TV, Books, Newspapers...



Real-world influence of Twitter

- ◎ Political campaigns
 - ◎ Marketing campaigns + promotions
 - ◎ Stock markets
 - ◎ Journalism: TV, Books, Newspapers...
 - ◎ Customer satisfaction
- 

Real-world influence of Twitter

- ◎ Political campaigns
- ◎ Marketing campaigns + promotions
- ◎ Stock markets
- ◎ Journalism: TV, Books, Newspapers...
- ◎ Customer satisfaction
- ◎ Awareness programs



Real-world influence of Twitter

- ◎ Political campaigns
- ◎ Marketing campaigns + promotions
- ◎ Stock markets
- ◎ Journalism: TV, Books, Newspapers...
- ◎ Customer satisfaction
- ◎ Awareness programs

A strong incentive to manipulate tweets



The problem

The problem

© No one knows if tweets can be trusted



The problem

- ◎ No one knows if tweets can be trusted
- ◎ Not even Twitter themselves
 - Researchers from Twitter

The problem

- ◎ No one knows if tweets can be trusted
- ◎ Not even Twitter themselves
 - Researchers from Twitter
 - Discussion with Vigiglobe

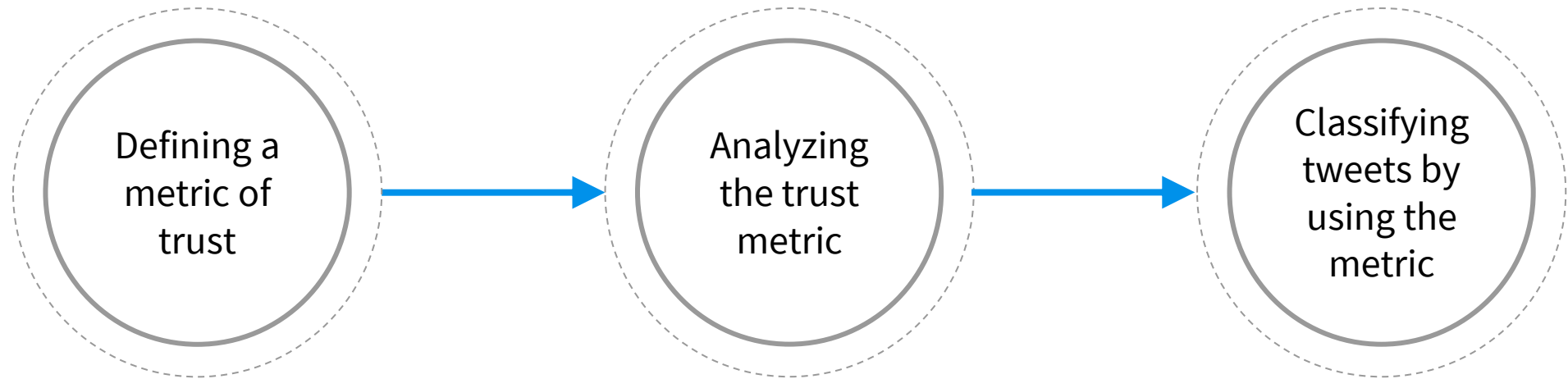


The problem

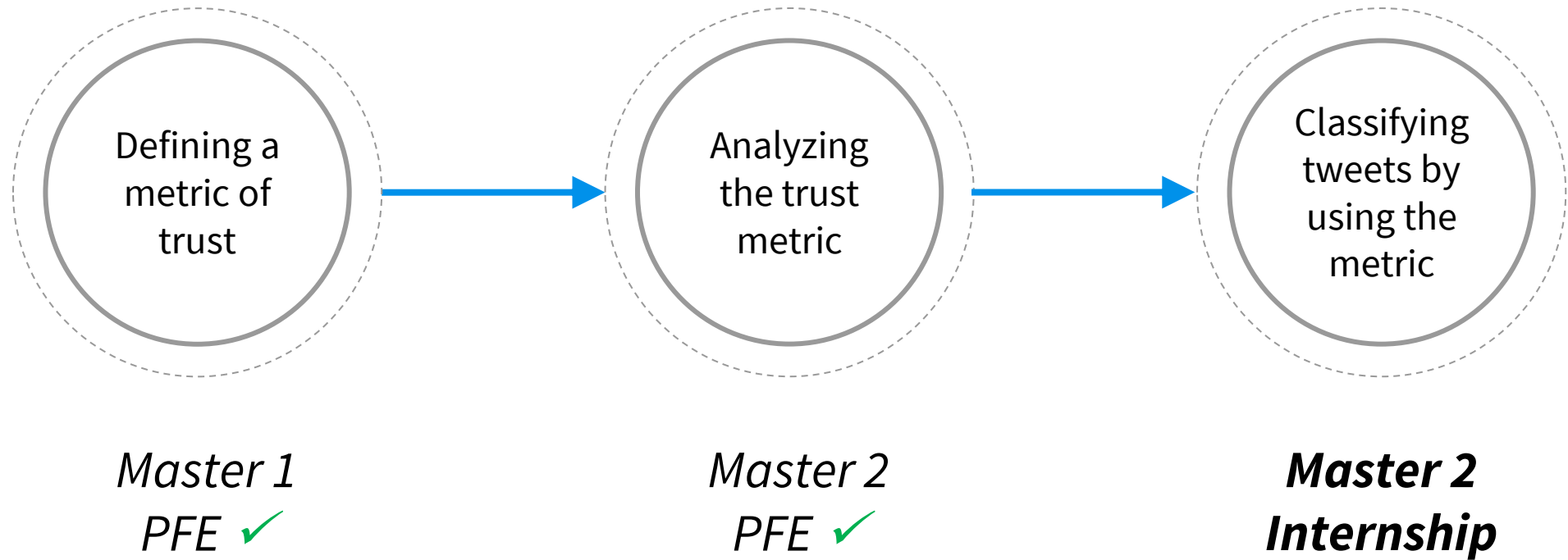
- ◎ No one knows if tweets can be trusted
- ◎ Not even Twitter themselves
 - Researchers from Twitter
 - Discussion with Vigiglobe
- ◎ Goal: Robust, **on-the-fly** spam detection



The workflow



The workflow



Outline

- ◎ Background, problem statement, workflow
- ◎ **Definition of our metric of trust**
- ◎ Spam detection methodology
- ◎ Testing our method
- ◎ Conclusion: Next Steps



Do we need a trust metric?

Do we need a trust metric?

◎ Twitter has manually verified ~ 113 K users

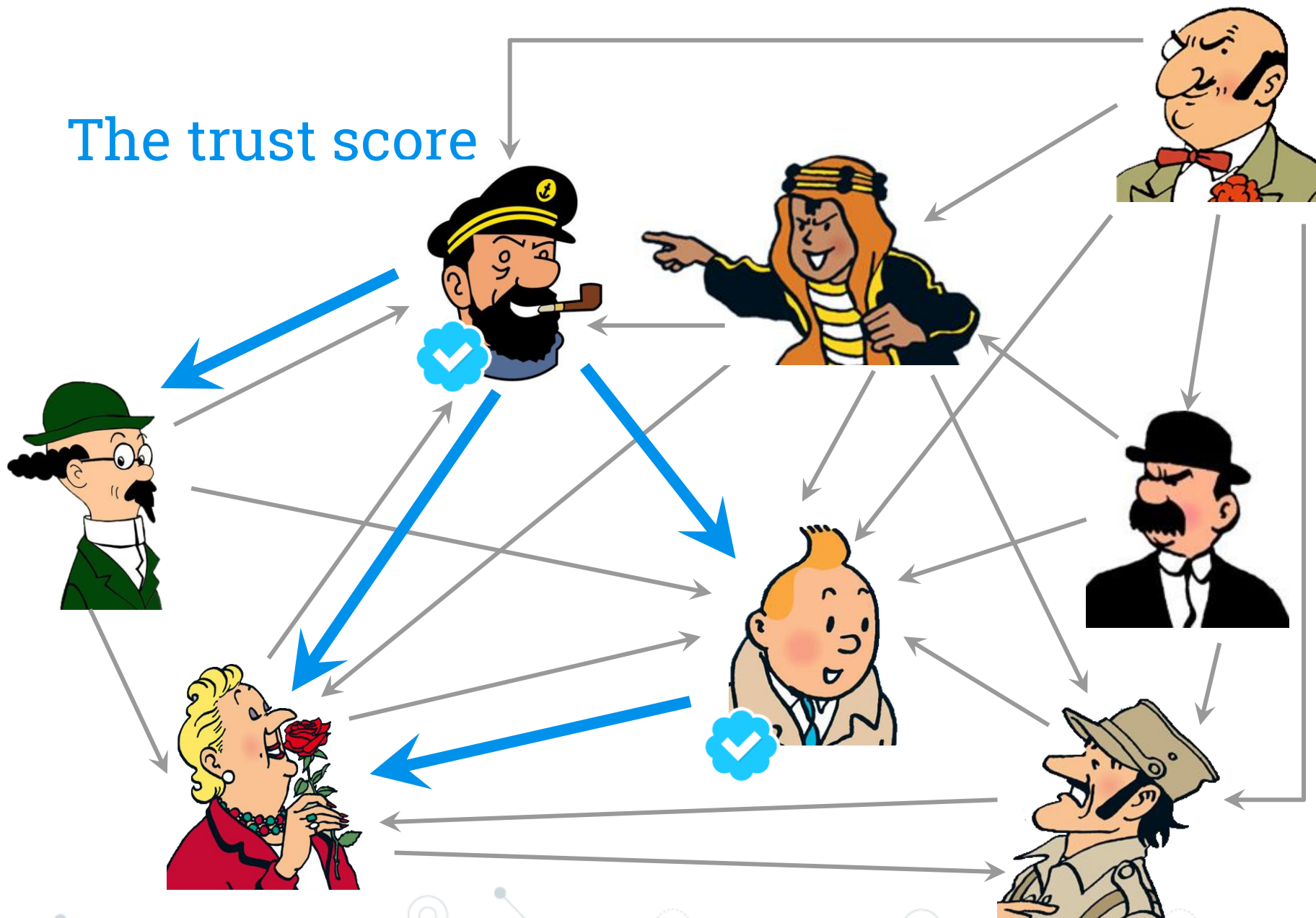


Do we need a trust metric?

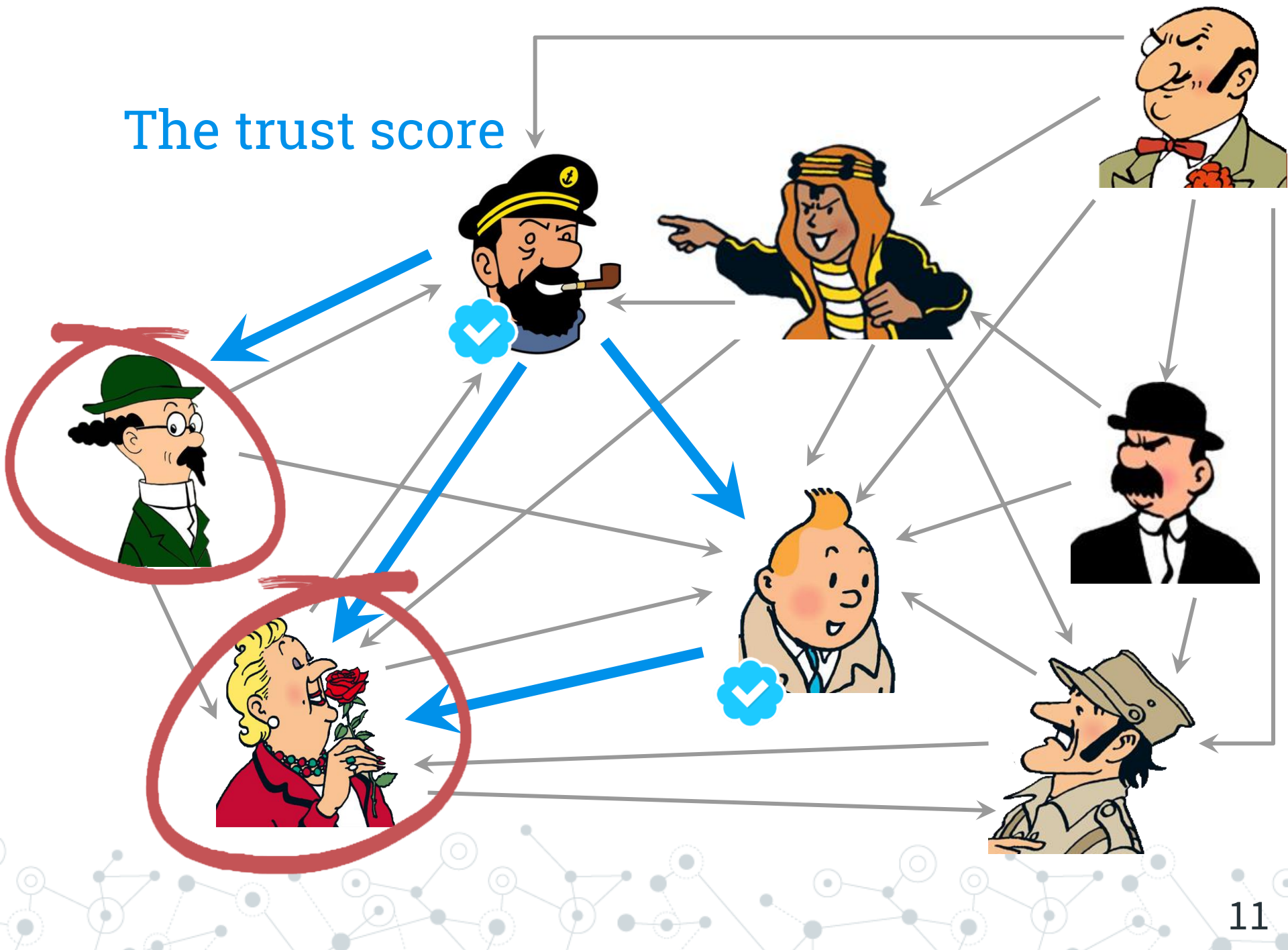
- ◎ Twitter has manually verified ~ 113 K users
- ◎ But **99.99 %** users are not verified



The trust score



The trust score



Outline

- ◎ Background, problem statement, workflow
- ◎ Definition of our metric of trust
- ◎ **Spam detection methodology**
- ◎ Testing our method
- ◎ Conclusion: Next Steps

Retweet chain



Retweet chain



Tintin, Reporter

@TintinIsForever



Follow

Tintin visited twice, but Hergé really ought to have done a full 'Tintin en Inde' (Tintin in India). [#TintinNostalgia](#)



Retweet chain

Creator of the tweet



Retweet chain

Creator of the tweet



Retweeters

The power of retweets



The power of retweets

- ◎ Non-repudiation: Public statement of one's *approval* of the content



The power of retweets

- ◎ Non-repudiation: Public statement of one's *approval* of the content
- ◎ Not duplication: Gives credit to the original content publisher



The power of retweets

- ◎ Non-repudiation: Public statement of one's *approval* of the content
- ◎ Not duplication: Gives credit to the original content publisher
- ◎ Long retweet chain = High visibility = Affects a lot of people

The power of retweets

- ◎ Non-repudiation: Public statement of one's *approval* of the content
- ◎ Not duplication: Gives credit to the original content publisher
- ◎ Long retweet chain = High visibility = Affects a lot of people
- ◎ Public opinion: Retweets influence trends

Spam detection method: Quality of retweets



Spam detection method: Quality of retweets

Trusted users in the retweet chain indicates
authenticity of the tweet



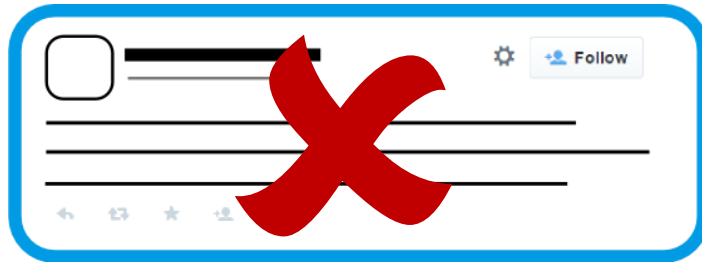
Spam detection method: Quality of retweets

Trusted users in the retweet chain indicates authenticity of the tweet



Spam detection method: Quality of retweets

Trusted users in the retweet chain indicates authenticity of the tweet





How is it robust and on the fly?



How is it robust and on the fly?

◎ Easy to send many tweets

How is it robust and on the fly?

- ◎ Easy to send many tweets
- ◎ Difficult to change the follow-relationship

How is it robust and on the fly?

- ◎ Easy to send many tweets
- ◎ Difficult to change the follow-relationship
- ◎ If we have the tweet, we can obtain the list of retweets, i.e. the retweet chain

Outline

- ◎ Background, problem statement, workflow
- ◎ Definition of our metric of trust
- ◎ Spam detection methodology
- ◎ **Testing our method**
- ◎ Conclusion: Next Steps

Testing our method of spam detection



Testing our method of spam detection

◎ No test set



Testing our method of spam detection

- ◎ No test set
- ◎ Manual verification too slow



Testing our method of spam detection

- ◎ No test set
- ◎ Manual verification too slow
- ◎ Need other methods



Testing our method of spam detection

- ◎ No test set
- ◎ Manual verification too slow
- ◎ Need other methods
 1. Suspicious keywords

Testing our method of spam detection

- ◎ No test set
- ◎ Manual verification too slow
- ◎ Need other methods
 1. Suspicious keywords
 2. Periodic tweets

Testing our method of spam detection

- ◎ No test set
- ◎ Manual verification too slow
- ◎ Need other methods
 1. Suspicious keywords
 2. Periodic tweets
 3. Content copying

Method 1: Keyword analysis



Method 1: Keyword analysis

- ◎ Collection of 27M topic-related tweets



Method 1: Keyword analysis

- ◎ Collection of 27M topic-related tweets
- ◎ Unrelated/derogatory keywords = spam?



- Collection of 27M topic-related tweets
- Unrelated/derogatory keywords = spam?

[illegible]

Method 1: Keyword analysis

- ◎ Collection of 27M topic-related tweets
- ◎ Unrelated/derogatory keywords = spam?

con sin hot bra fuck lol giveaway mom kill sale girl hack
upgrade bf prom hole exe sex fuckin leak gratis fucking
suck followers torrent cheap tops fucked password girls
ninja retweets kick male killing dude bitch recent kills
gay baby nights hackers cute repair discount pirates
rumor teen sexy porn followme fake death finger
giveaways wife playroom dick died hiring subscribe
multiplayer rear spy midnight dumb upgrades pissed
peek freak killer webcam shirt sponsor models cheapest
wallpaper installation

Method 1: Keyword analysis

- ◎ Collection of 27M topic-related tweets
- ◎ Unrelated/derogatory keywords = spam?

con sin hot bra fuck lol giveaway mom kill sale girl hack
upgrade bf prom hole exe sex fuckin leak gratis fucking
suck followers torrent cheap tops fucked password girls
ninja retweets kick male killing dude bitch recent kills
gay baby nights hackers cute repair discount pirates
rumor teen sexy **porn** followme fake death finger
giveaways wife playroom dick died hiring subscribe
multiplayer rear spy midnight dumb upgrades pissed
peek freak killer webcam shirt sponsor models cheapest
wallpaper installation

Method 1: Keyword analysis

- ◎ Collection of 27M topic-related tweets
- ◎ Unrelated/derogatory keywords = spam?



Method 1: Keyword analysis

- ◎ Collection of 27M topic-related tweets
- ◎ Unrelated/derogatory keywords = spam?



Method 1: Keyword analysis

- ◎ Collection of 27M topic-related tweets
- ◎ Unrelated/derogatory keywords = spam?

Unrelated/suspicious keyword detection is *not* a good way to detect spam

Method 2: Finding periodic tweets



Method 2: Finding periodic tweets

- ◎ Twitter bots often tweet periodically



Method 2: Finding periodic tweets

- ◎ Twitter bots often tweet periodically
- ◎ Difficult to detect periods in a large collection of tweets



Method 3: Replication of tweet content



Method 3: Replication of tweet content

- ◎ Some tweets have the same content



Method 3: Replication of tweet content

- ◎ Some tweets have the same content
- ◎ Same content → **Spam property**

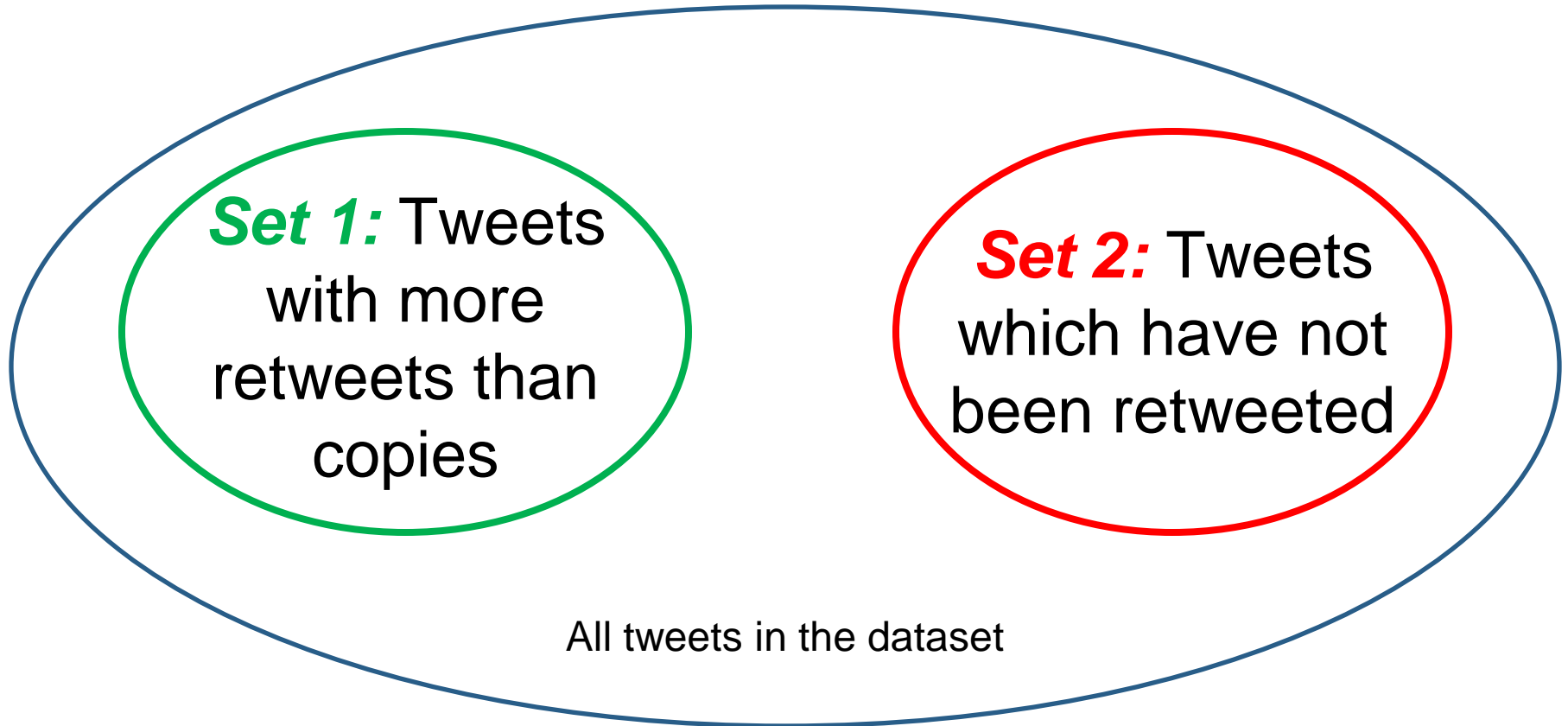


Method 3: Replication of tweet content

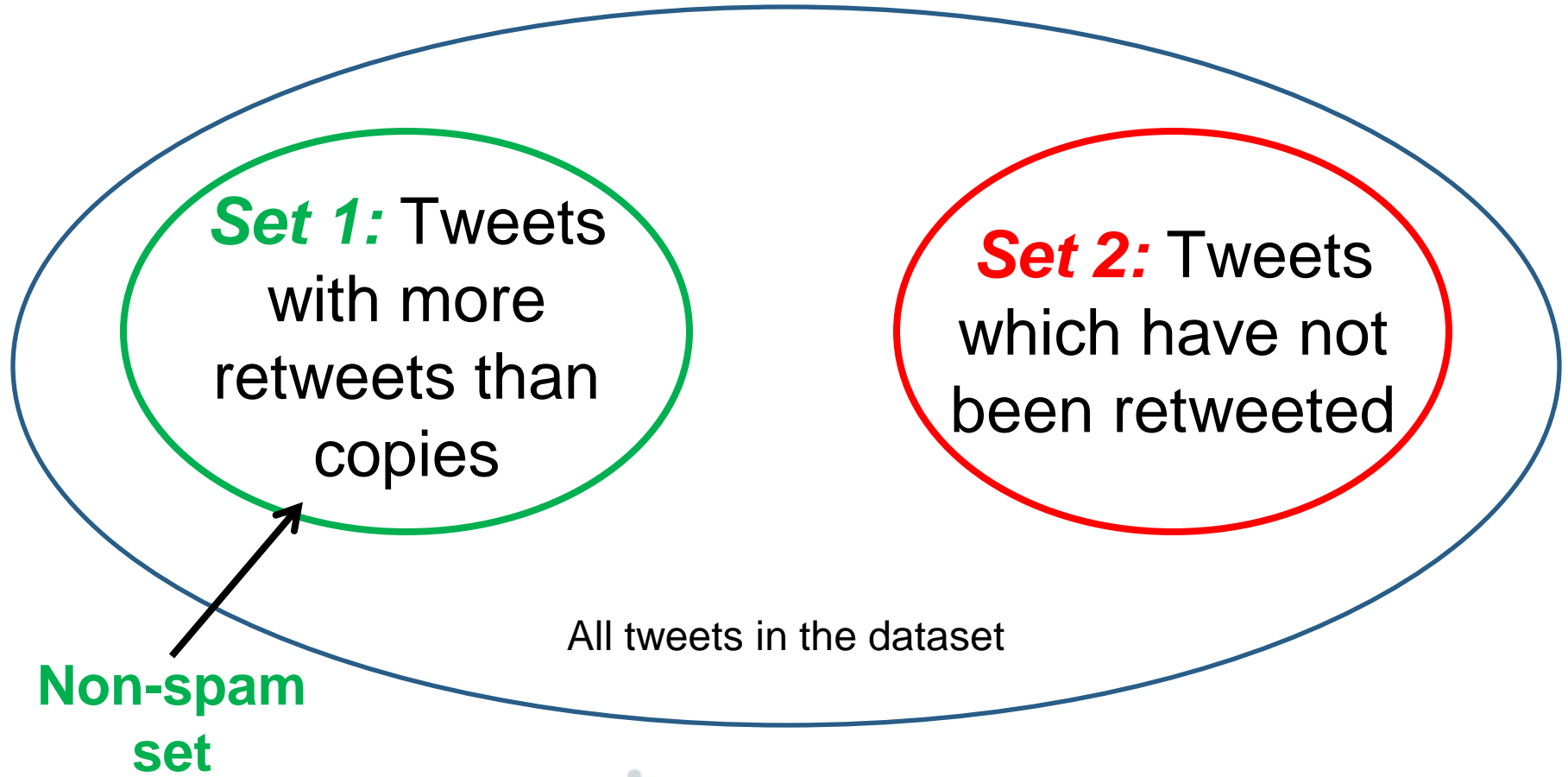
- ◎ Some tweets have the same content
- ◎ Same content → **Spam property**
- ◎ Retweets → **Non-spam property**



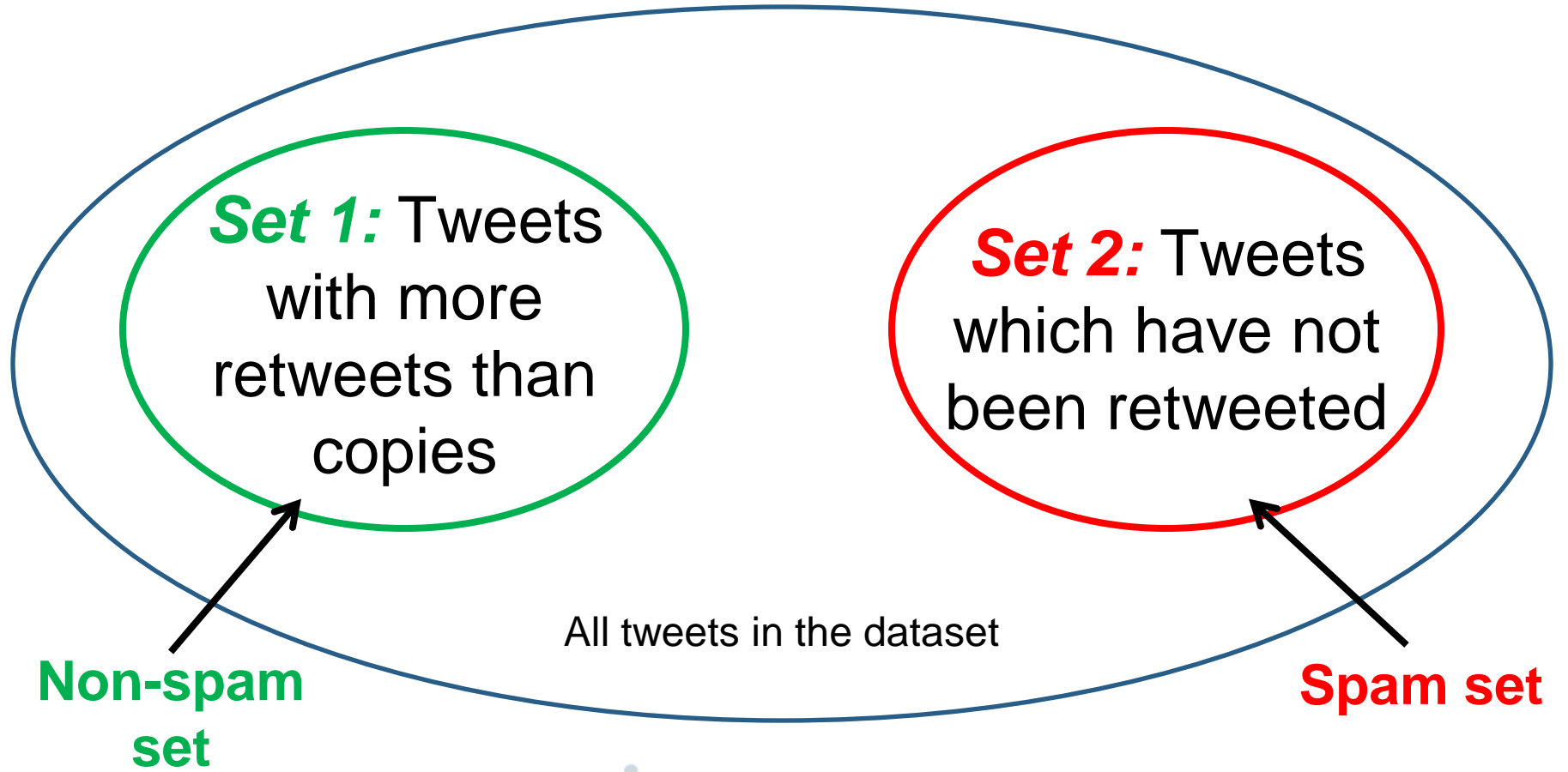
Method 3: Replication of tweet content



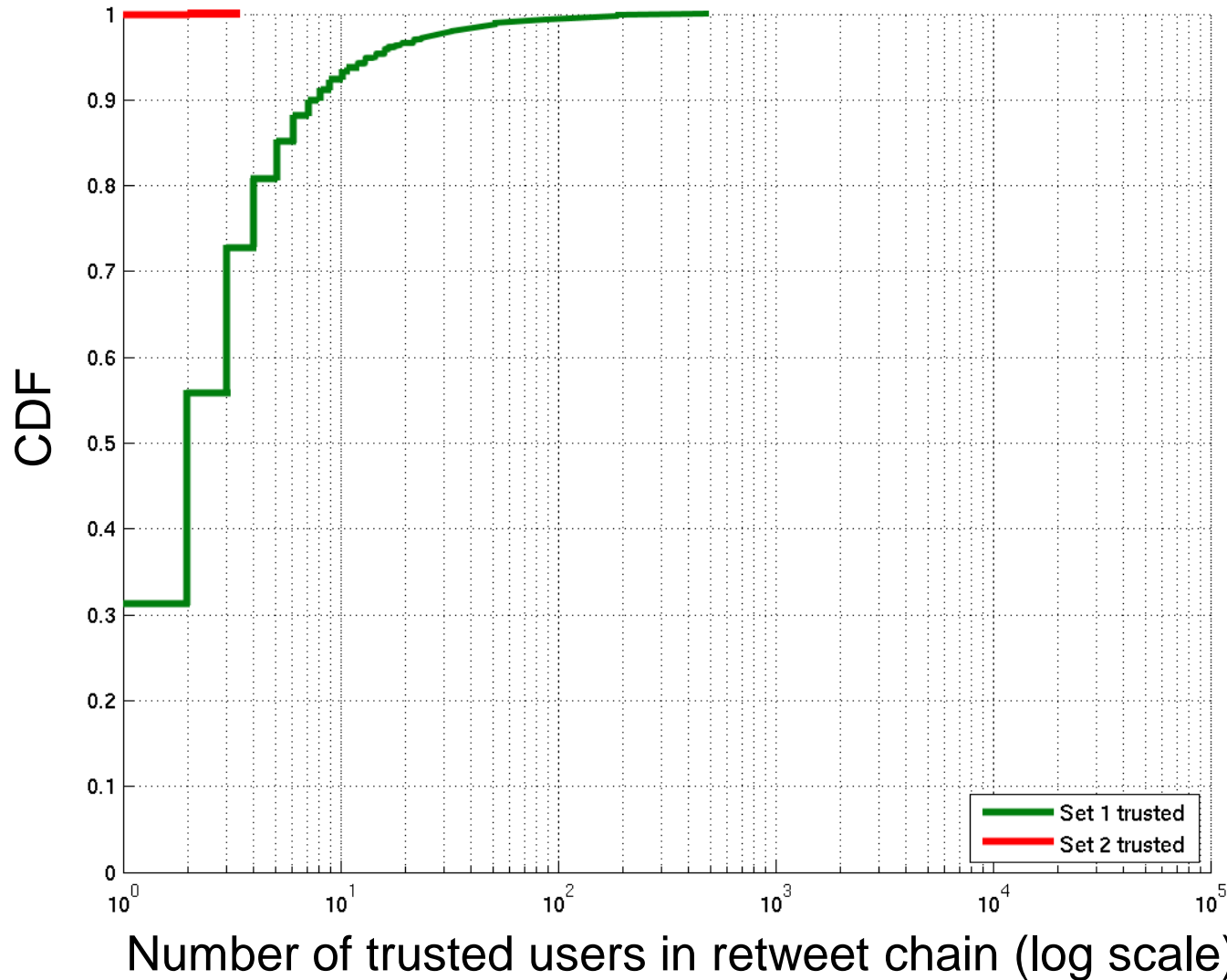
Method 3: Replication of tweet content



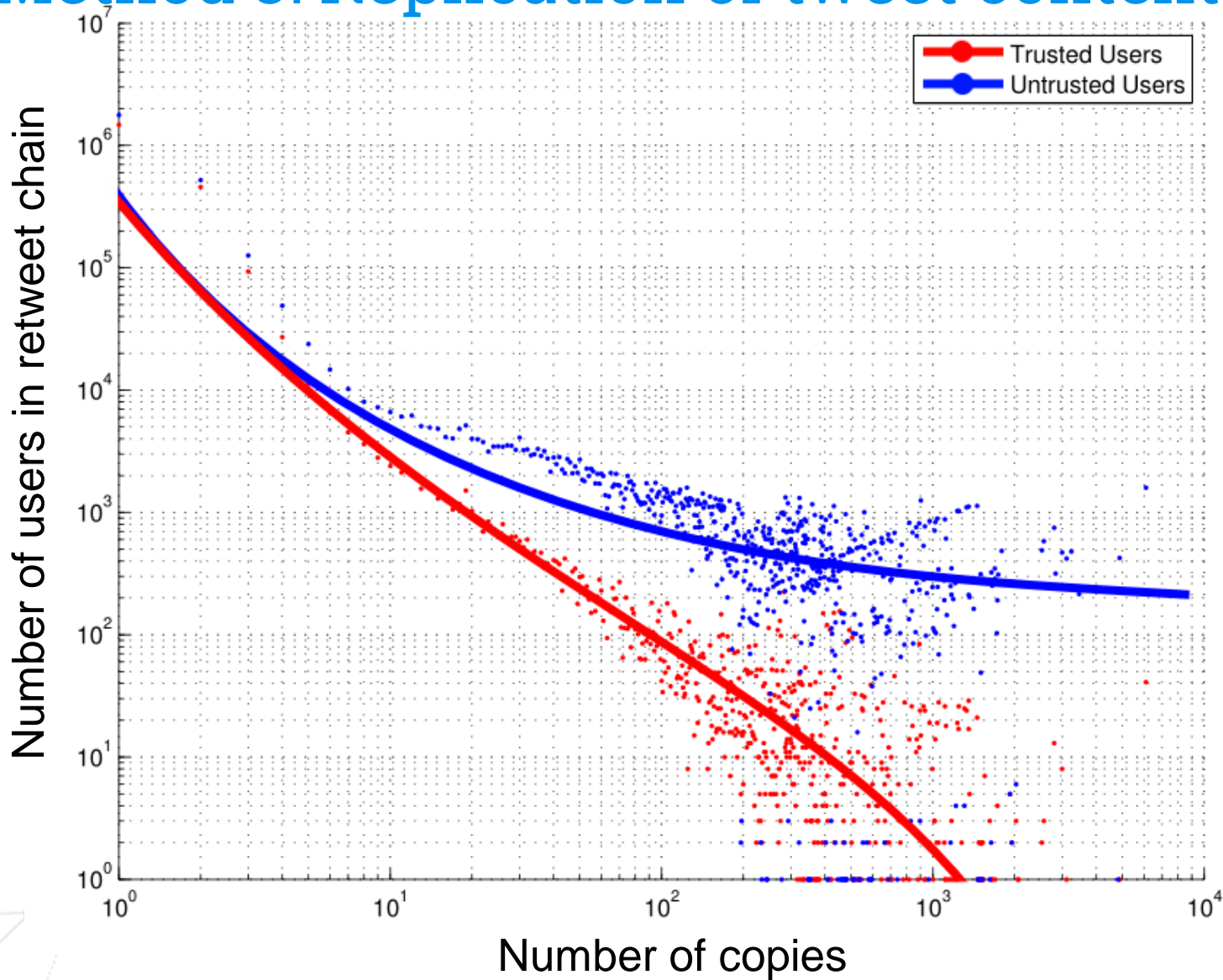
Method 3: Replication of tweet content



Method 3: Replication of tweet content



Method 3: Replication of tweet content





Outline

- ◎ Background, problem statement, workflow
- ◎ Definition of our metric of trust
- ◎ Spam detection methodology
- ◎ Testing our method
- ◎ **Conclusion: Next Steps**



Next steps: Plan for the next two months



Next steps: Plan for the next two months

- ◎ Testing our method in other datasets



Next steps: Plan for the next two months

- ◎ Testing our method in other datasets
- ◎ Correlation with other methods

Next steps: Plan for the next two months

- ◎ Testing our method in other datasets
- ◎ Correlation with other methods
- ◎ Spam detection as a service/API

Conclusion



Conclusion

◎ **On-the-fly** spam detection



Conclusion

- ◎ **On-the-fly** spam detection
- ◎ Help prevent manipulation of public opinion on Twitter



Conclusion

- ◎ **On-the-fly** spam detection
- ◎ Help prevent manipulation of public opinion on Twitter

Making social networks safer and more authentic





Thank you!

How to find spam on Twitter?

Mourjo Sen

Under the guidance of

Arnaud Legout, Maksym Gabielkov