Machine Learning and Cybercrime

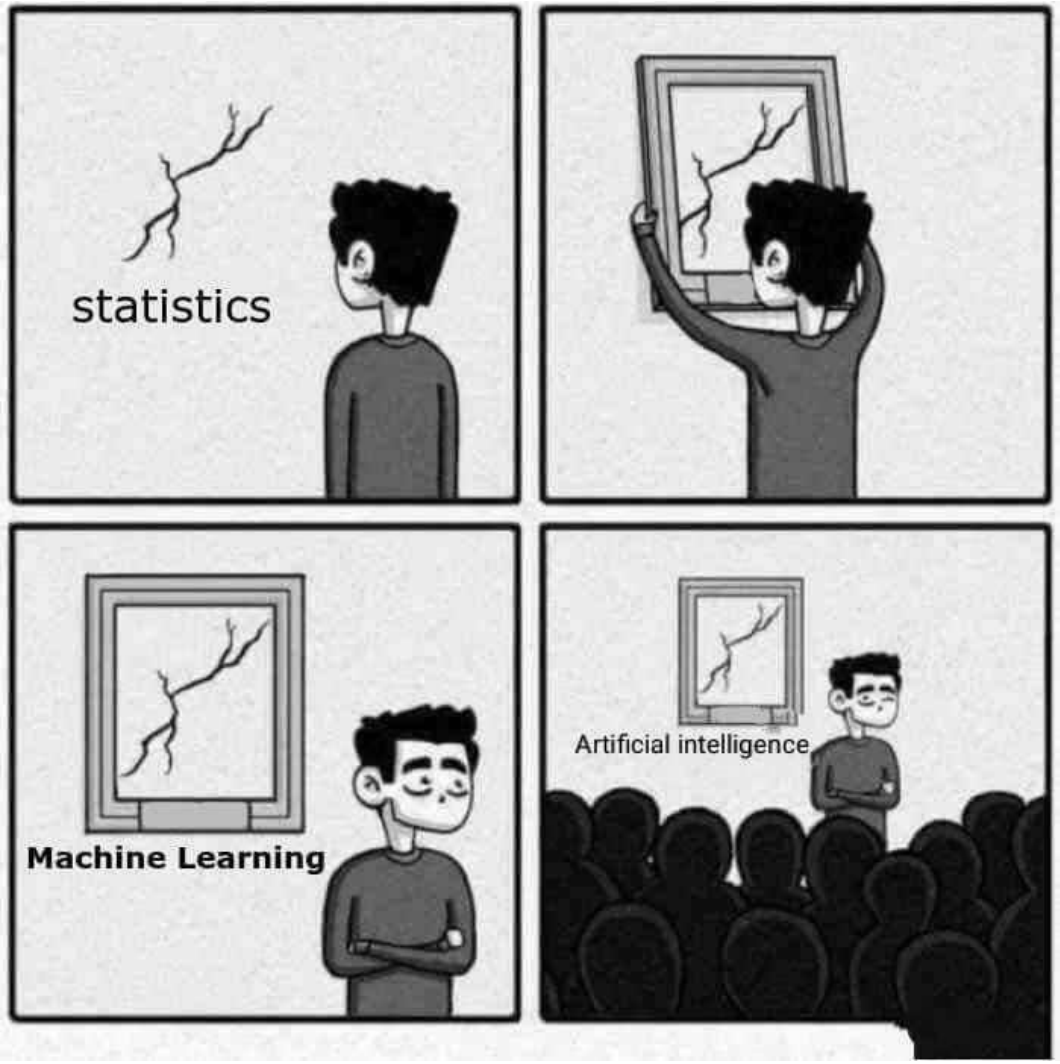
Rebekah Overdorf

Panoramix

KU LEUVEN RESEARCH & DEVELOPMENT

What is Machine Learning?

Machine learning uses statistical techniques to give computers the ability to "learn" with data, without being explicitly programmed.



Traditional Programming

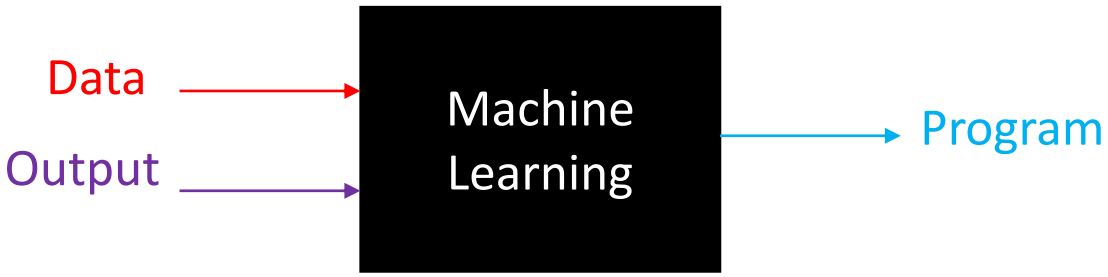
- Know something about the task
- Explicitly tell the machine how to complete it
- Test the model on something you know
- Deploy your program!

Traditional Programming Example: Hate Speech Detection

- Know something about the task
- Explicitly tell the machine how to complete it
- Test the model on something you know
- Deploy your program!

- Hate speech usually contains the words in this list I made: list.txt
- Write a program that looks for these words
- Test the model on some hate speech that I found. “Wow it works on a bunch of these!”
- Deploy!

Machine Learning?



- Data Mining: Get data that you know something about
- Train the Model: “Teach” the machine about that data
- Test the model on something you know
- Deploy the program!

Machine Learning Example: Hate Speech Detection

- Data Mining: Get data that you know something about
- Train the Model: “Teach” the machine about that data
- Test the model on something you know
- Deploy the program!

- I have 2 data sets, one of hate speech and one of not hate speech
- I throw all of this data at a “classifier” that uses statistics to figure out what is hate speech and what isn’t.
- Test the model on some more hate speech that I found. “Wow it works on a bunch of these!”
- Deploy!

This “machine learning” is super useful

- For understanding cybercrime to
  - **Predict their communications**
    - **Overdorf et. al. 2018 (arxiv)**
  - Link accounts using writing style
    - Afroz et. al. 2012
  - Detect Phishing Sites and Emails
    - Xiang et. al. 2011, Basnet et. al. 2008
  - Identify items for sale/transactions
    - Portnoff et. al. 2017
- Identify Scams
  - McCoy et. al. 2016
- Identify Hate Speech
  - Davidson et. al. 2017
- Identify Child Porn
  - Peersman et. al. 2012, 2016
- Study their structure
  - **Garg et. al 2015**

private messenger, post status updates, manage your profile and so much more. If you already have an account, login here - otherwise create an account for free today!

Page 1 of 2 1 2 next

This topic is locked

PRIVATE INVESTIGATION SERVICE Dox Anyone Accurate Results Advanced Techniques BTC Accepted

StackFramed
Offline

StackFramed

previously Rigzorra

MEMBER

Posts:	25
Joined:	Jun 25, 2017
Reputation:	7
Likes:	6
Credits:	0
Leecher level:	43

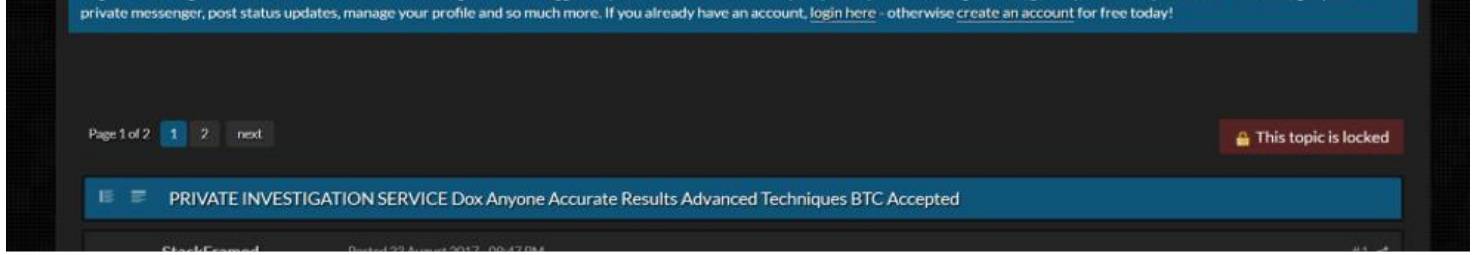
HALF YEAR REGISTERED

Posted 23 August 2017 - 09:47 PM #1

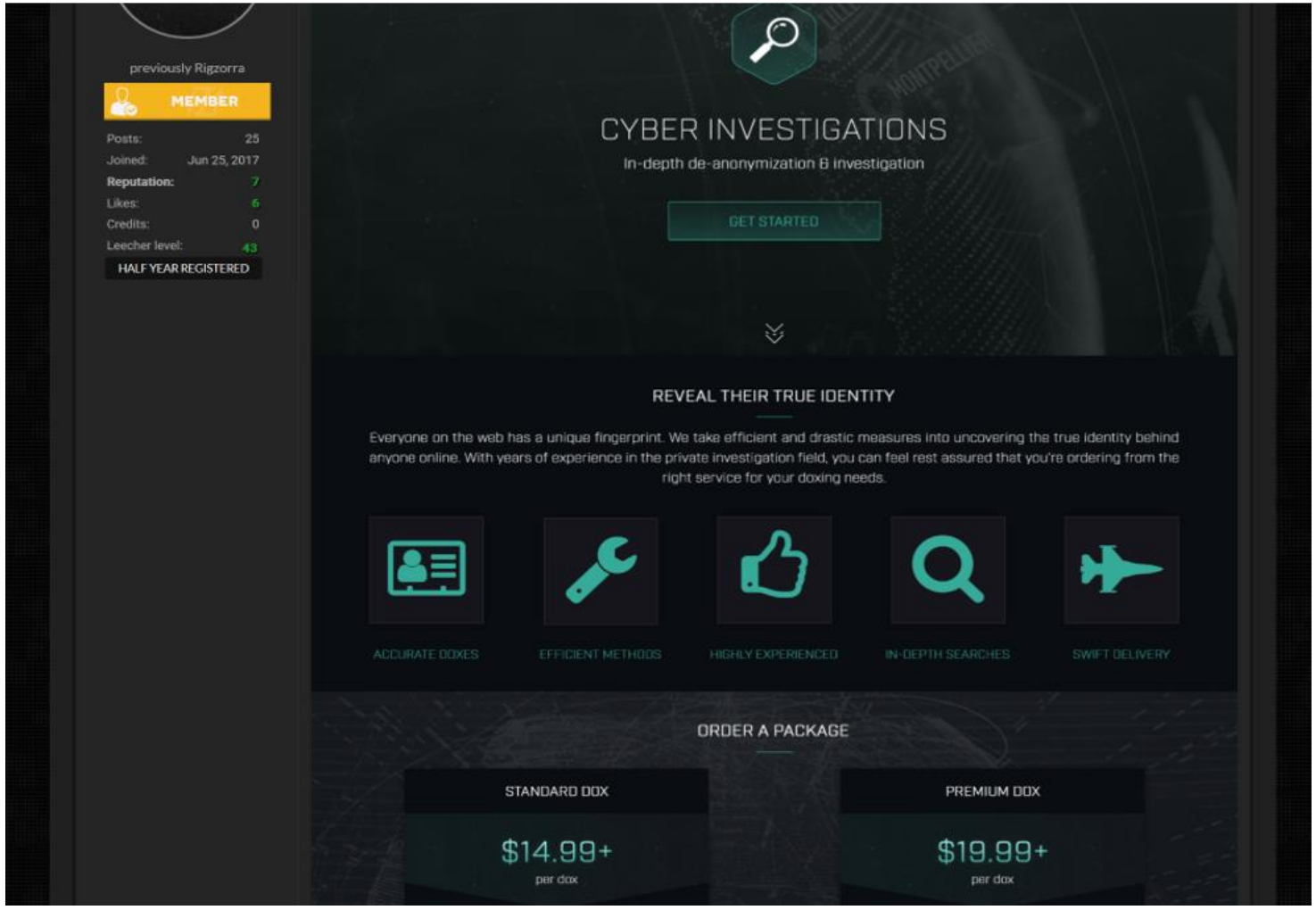
REVEAL THEIR TRUE IDENTITY

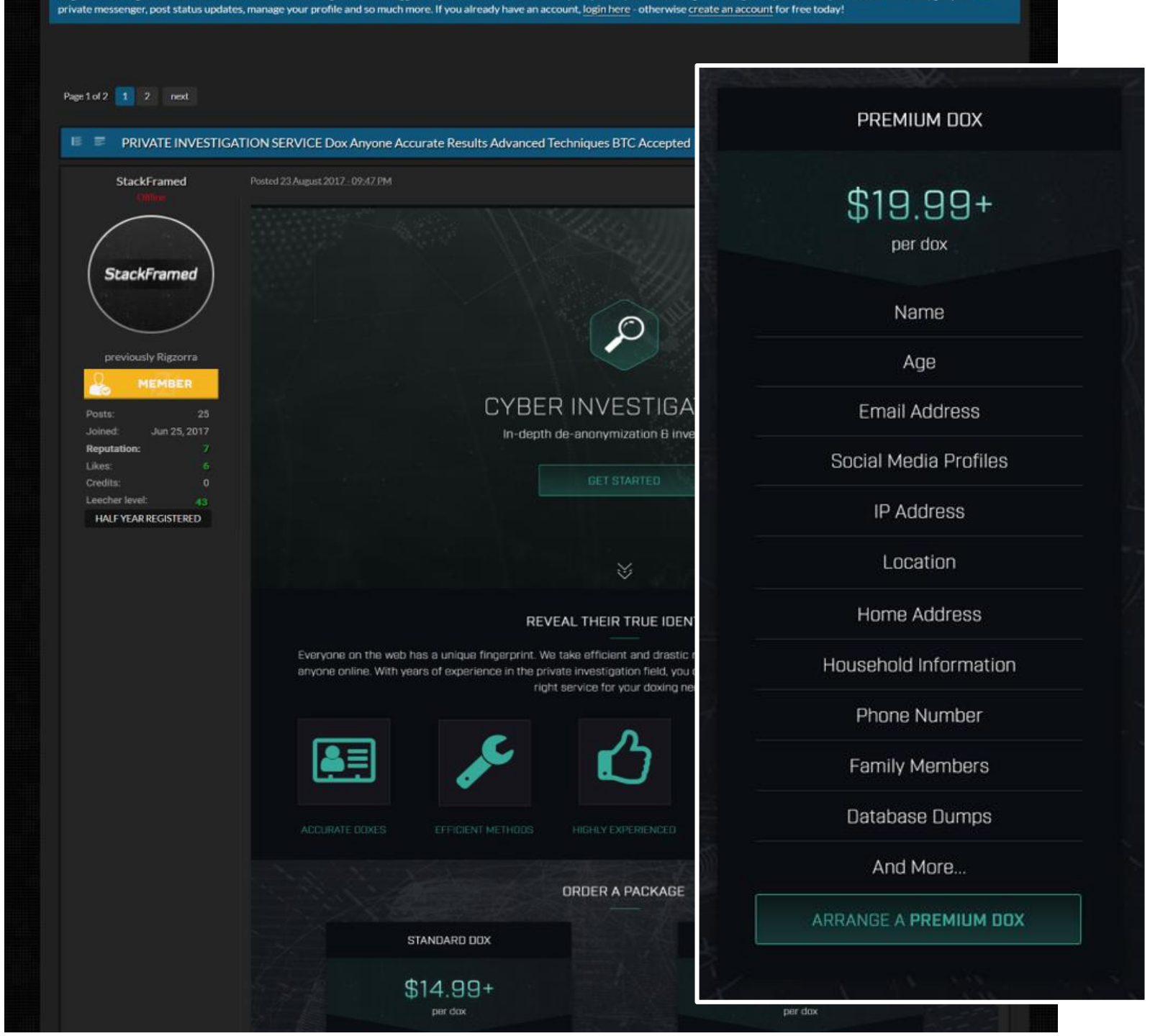
Everyone on the web has a unique fingerprint. We take efficient and drastic measures into uncovering the true identity behind anyone online. With years of experience in the private investigation field, you can feel rest assured that you're ordering from the right service for your doxing needs.

ORDER A PACKAGE



PRIVATE INVESTIGATION SERVICE Dox Anyone Accurate Results Advanced Techniques BTC Accepted




private messenger, post status updates, manage your profile and so much more. If you already have an account, login here - otherwise create an account for free today!
Page 1 of 2 1 2 next
PRIVATE INVESTIGATION SERVICE Dox Anyone Accurate Results Advanced Techniques BTC Accepted
StackFramed
Offline
Posted 23 August 2017 - 09:47 PM
StackFramed
previously Rigzorra
MEMBER
Posts: 25
Joined: Jun 25, 2017
Reputation: 7
Likes: 6
Credits: 0
Leecher level: 43
HALF YEAR REGISTERED
CYBER INVESTIGA
In-depth de-anonymization & inve
GET STARTED
REVEAL THEIR TRUE IDEN
Everyone on the web has a unique fingerprint. We take efficient and drastic
anyone online. With years of experience in the private investigation field, you
right service for your doxing ne
ACCURATE DOXES
EFFICIENT METHODS
HIGHLY EXPERIENCED
ORDER A PACKAGE
STANDARD DOX
$14.99+
per dox
PREMIUM DOX
$19.99+
per dox
Name
Age
Email Address
Social Media Profiles
IP Address
Location
Home Address
Household Information
Phone Number
Family Members
Database Dumps
And More...
ARRANGE A PREMIUM DOX
per dox

Public

Reply to this post

Private

Send message

Public

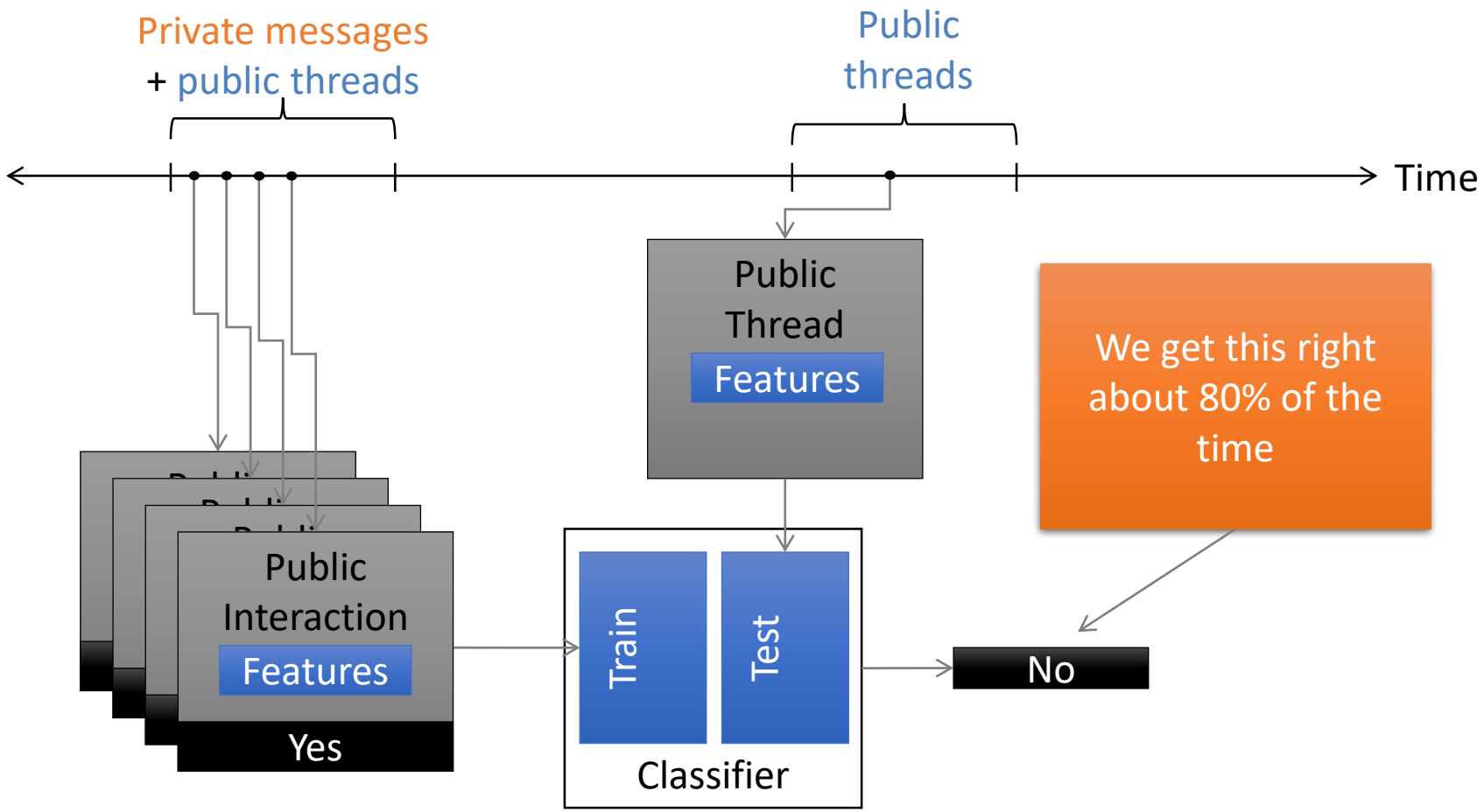
Reply to this post

We always have this

Private

Send message

We only have this when someone leaks data

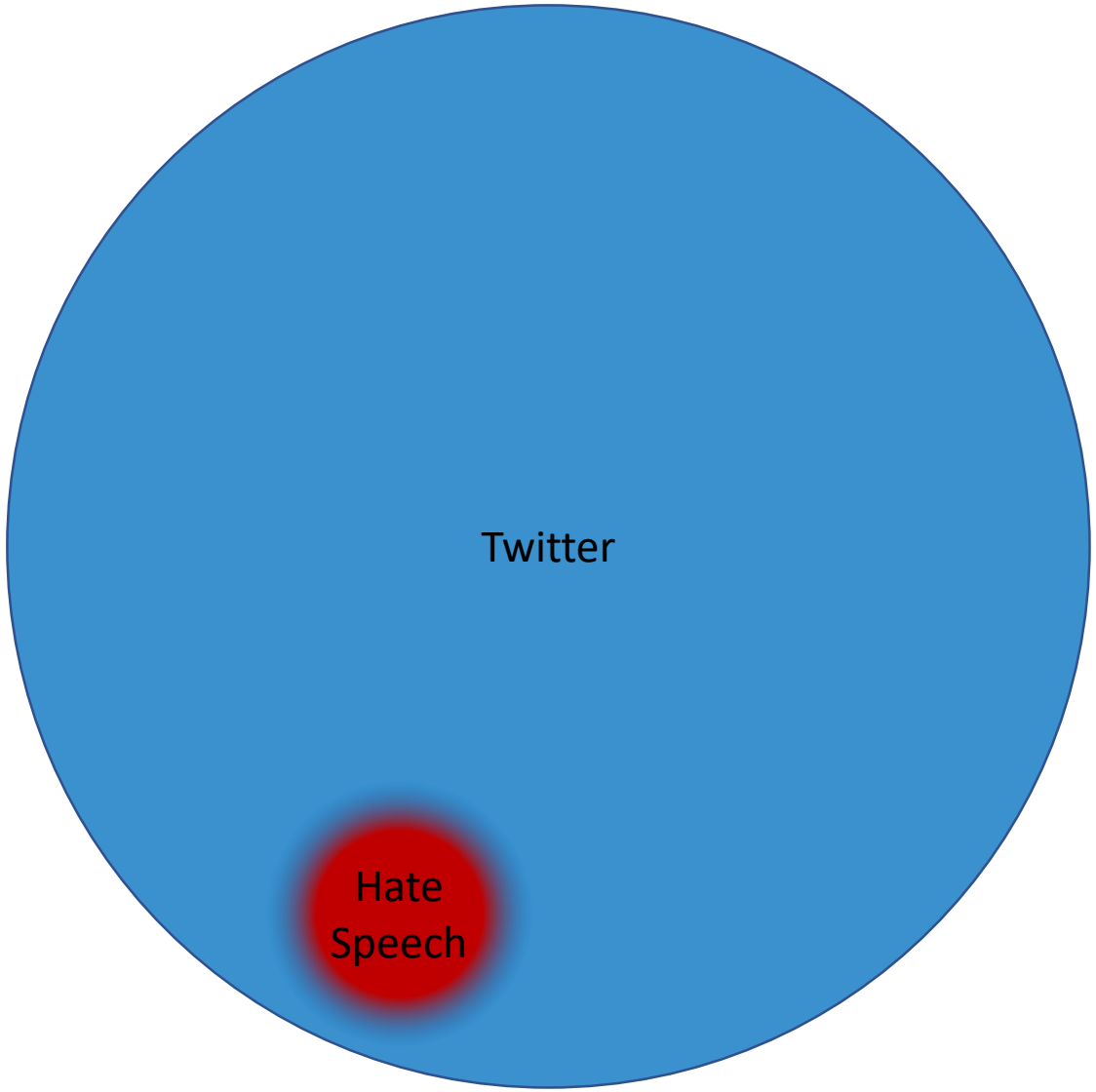

Private messages
+ public threads
Public
threads
Time
Public
Thread
Features
We get this right
about 80% of the
time
Public
Interaction
Features
Yes
Train
Test
Classifier
No

Potential Issues with deploying AI Systems

- ML often finds correlation, not causation
- Bias
- **Base Rate Fallacy**
- Privacy (e.g. membership attacks)
- Adversarial ML threats
- Result in antisocial and negative environmental outcomes
- Have adverse side effects
- Are built to only benefit a subset of users
- Externalize risks
- **Produce errors due to ds distributional shift**
- Result in systems that exploit states that fulfill the objective function, but do not complete the intended task
- **Distribute errors unfairly**
- Lack of transparency
- …

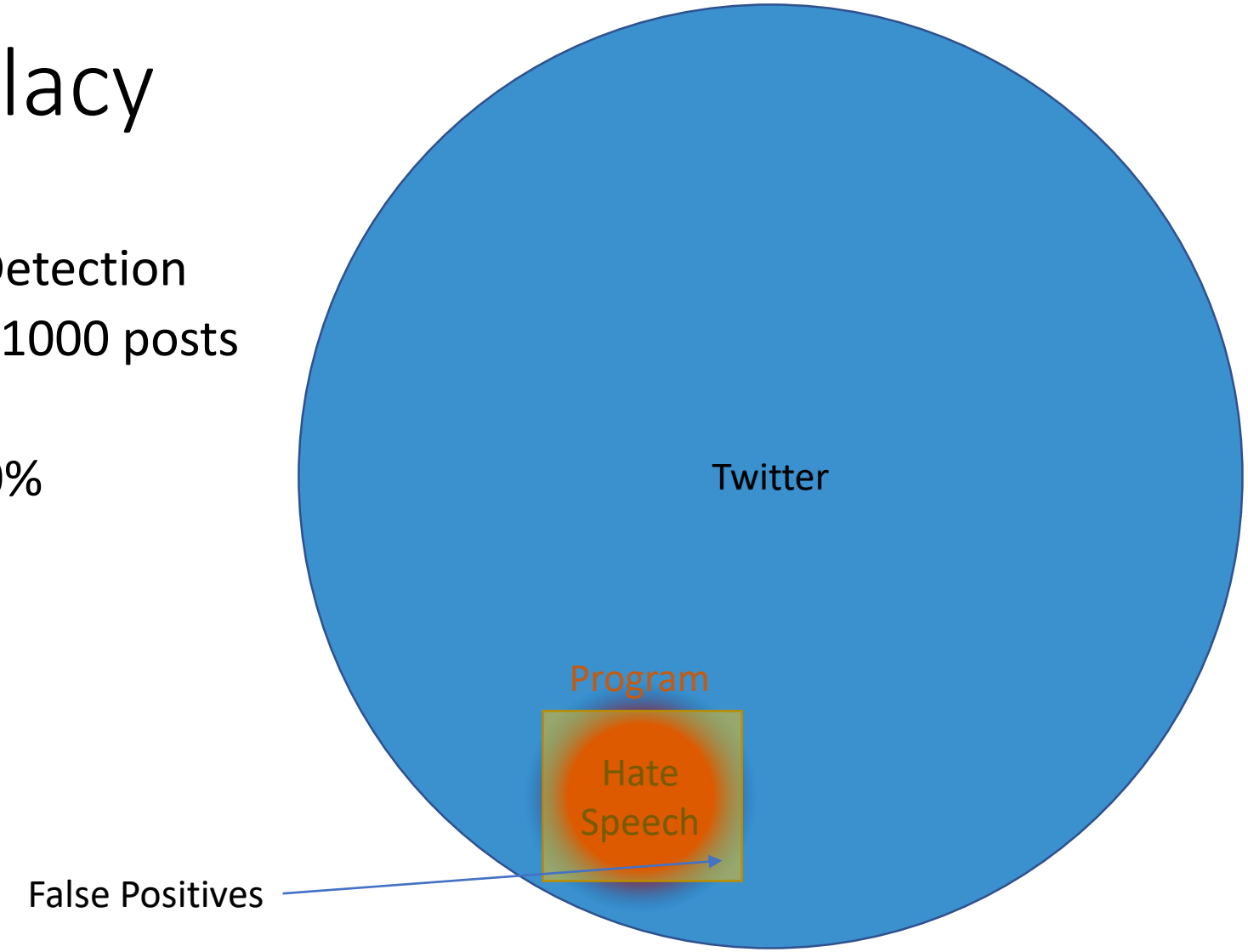
Base Rate Fallacy

- Example: Hate Speech Detection
- Hate Speech: 1 in every 1000 posts



Base Rate Fallacy

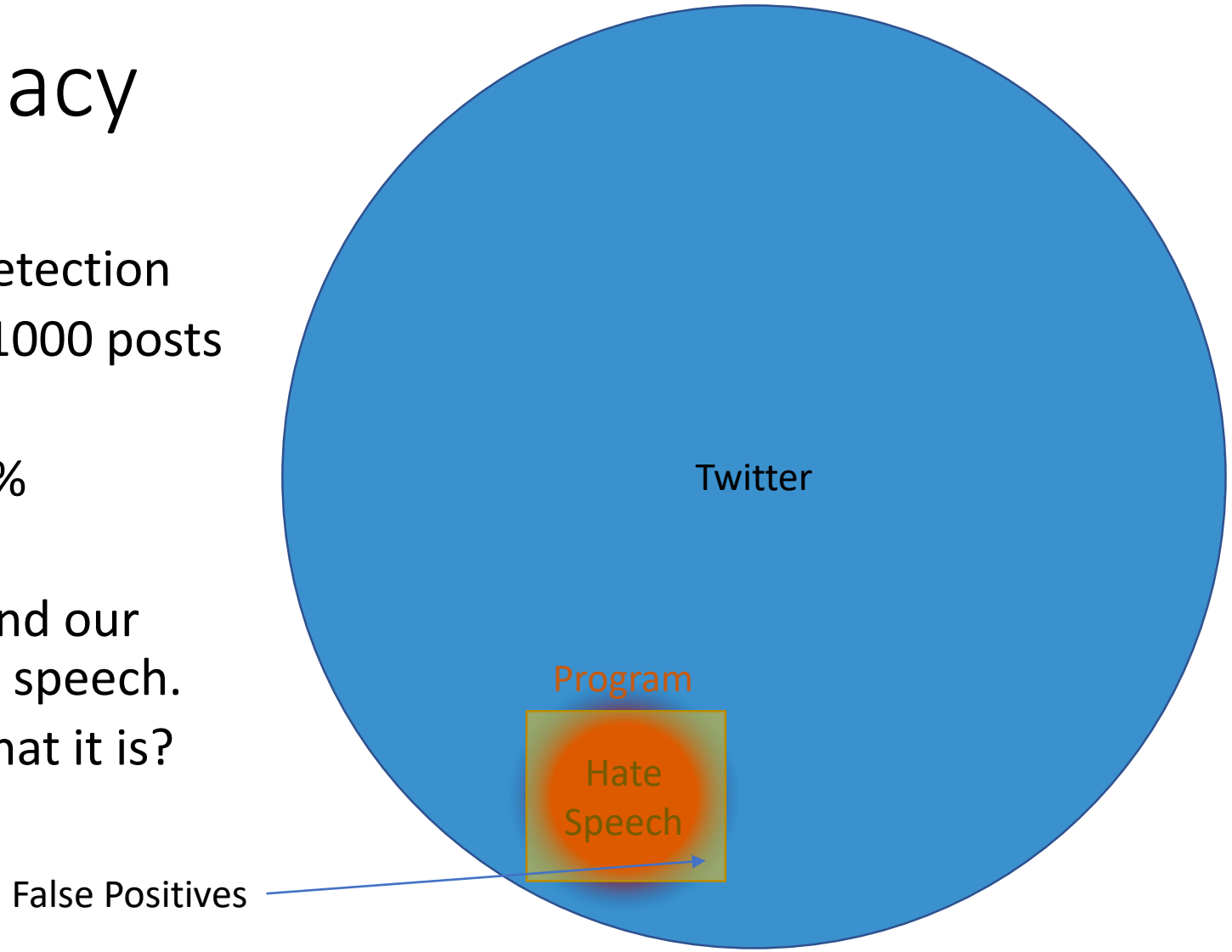
- Example: Hate Speech Detection
- Hate Speech: 1 in every 1000 posts
- False Positive rate of 5%
- True Positive rate of 100%



Base Rate Fallacy

- Example: Hate Speech Detection
- Hate Speech: 1 in every 1000 posts
- False Positive rate of 5%
- True Positive rate of 100%

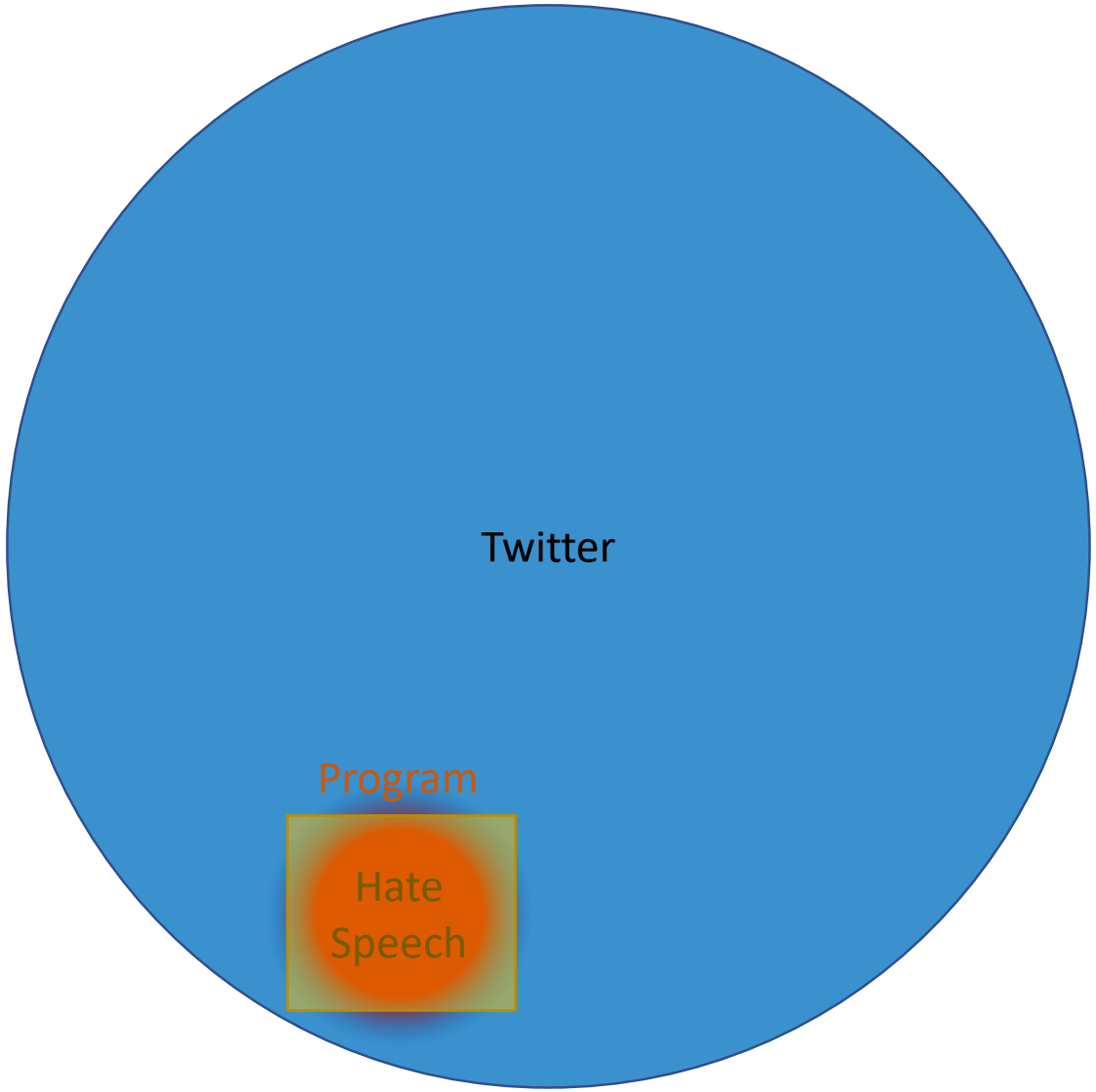
- A new tweet is posted, and our ML method says it's hate speech.
- What is the probability that it is?



Base Rate Fallacy

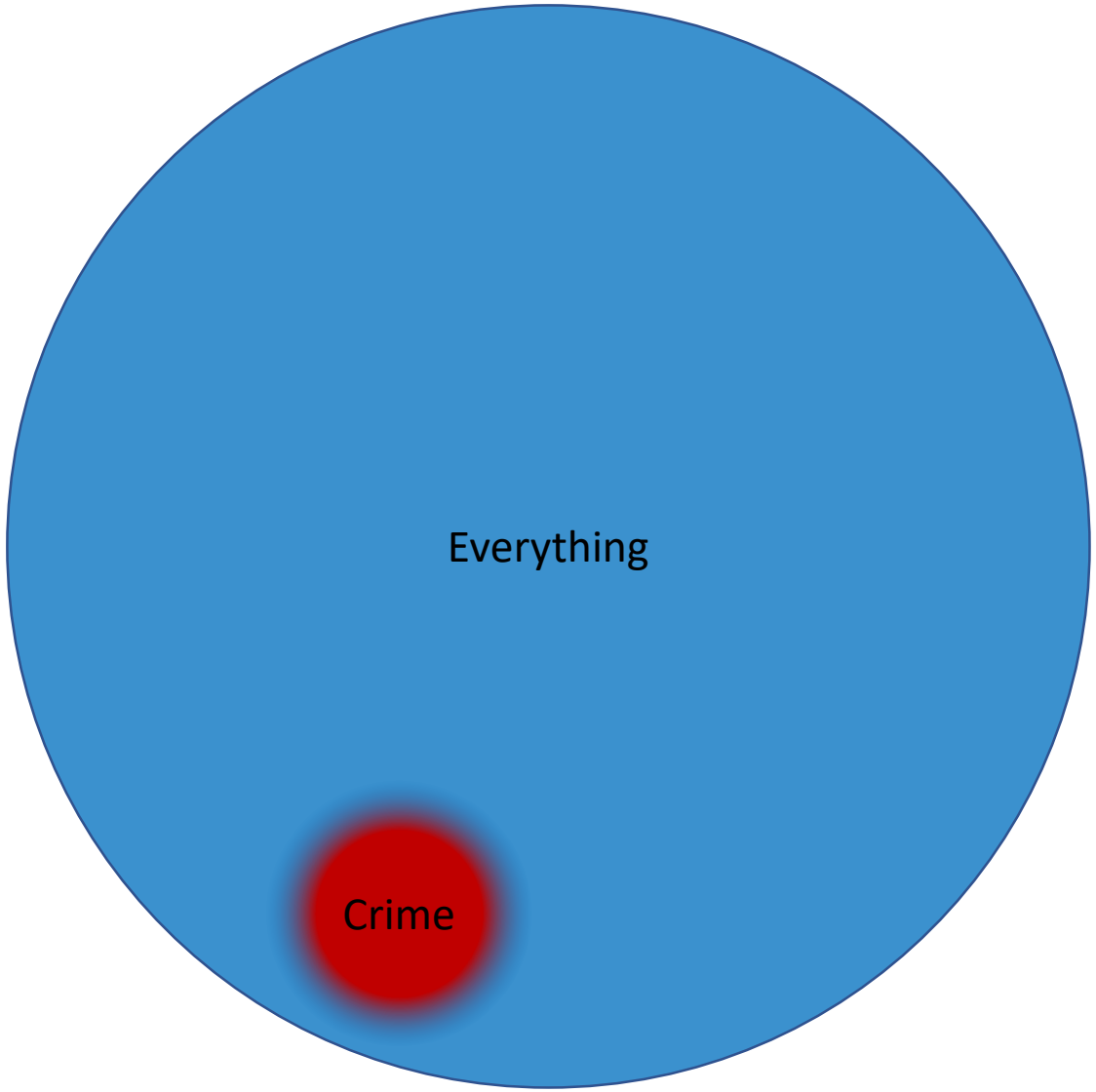
- Example: Hate Speech Detection
- Hate Speech: 1 in every 1000 posts
- False Positive rate of 5%
- True Positive rate of 100%

- A new tweet is posted, and our ML method says it's hate speech.
- What is the probability that it is?
  - ~2%
  - $p(hate|H) = \frac{p(H|hate)p(hate)}{p(H)}$
  - $p(hate|H) = \frac{1 * 0.001}{0.05095}$



Base Rate Fallacy

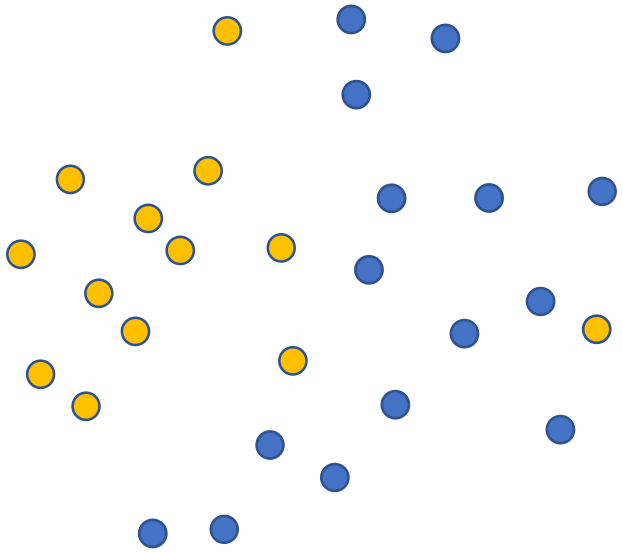
- ML can assist in decisions related to cybercrime, but commonly the base rate is skewed. **Most things aren't crime.**
- Adding a person to inspect the decisions is crucial.



Distribution of Errors

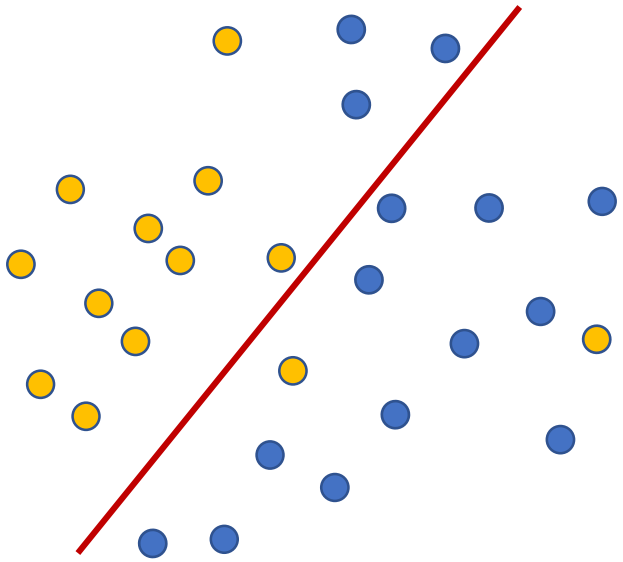
- Occurs when a all of the mistakes of the classifier are distributed to a subpopulation/group

Distribution of Errors



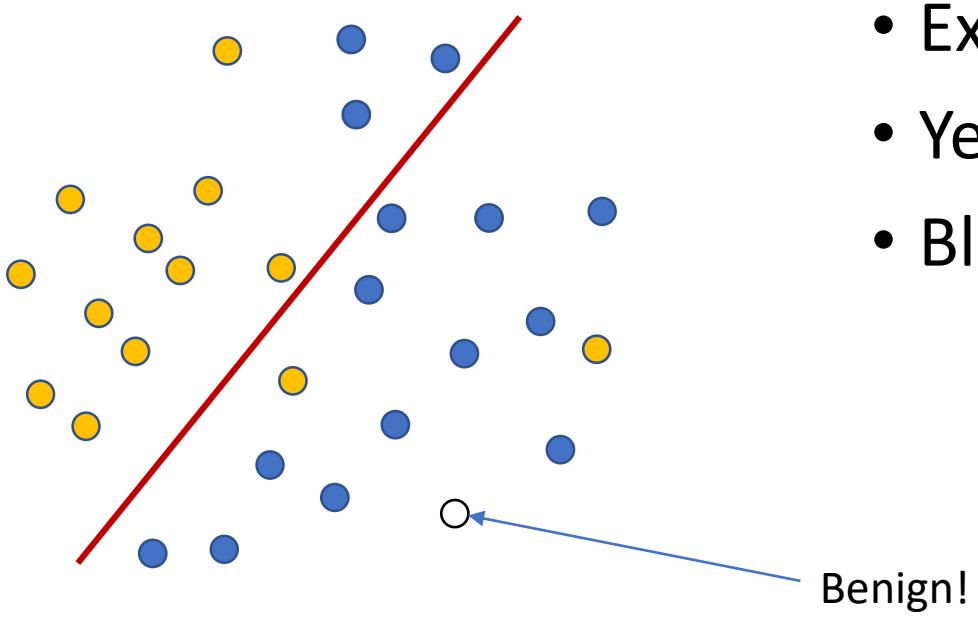
- Example: Email Phishing Detection
- Yellow = Yes phishing email
- Blue = Benign

Distribution of Errors



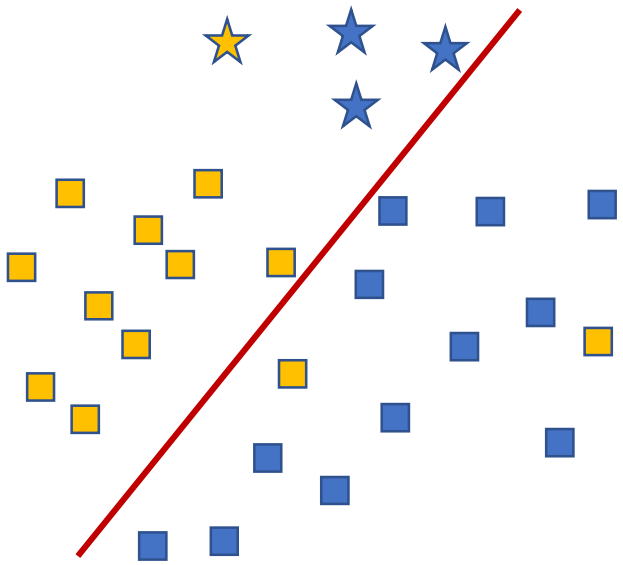
- Example: Email Phishing Detection
- Yellow = Yes phishing email
- Blue = Benign

Distribution of Errors



- Example: Email Phishing Detection
- Yellow = Yes phishing email
- Blue = Benign

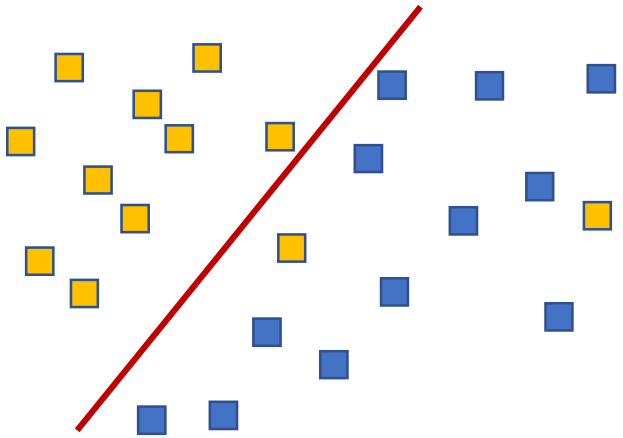
Distribution of Errors



- Example: Email Phishing Detection
- Yellow = Yes phishing email
- Blue = Benign
- Squares = Emails in English
- Stars = Emails in German

Distributional Shift

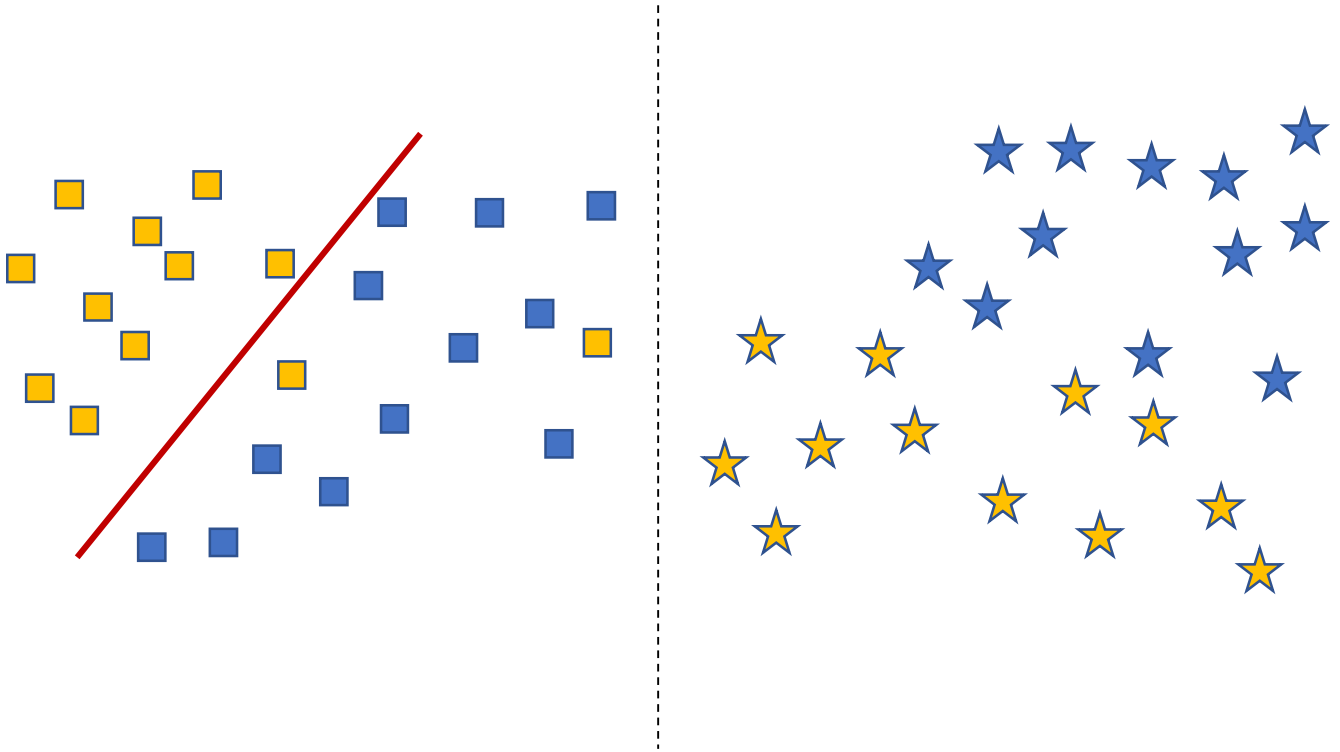
- Occurs when a classifier is trained in one area and deployed in another.



- Example: Bot or Not?
- Squares = Bots in the US
- Stars = Bots in Central Asia

Distributional Shift

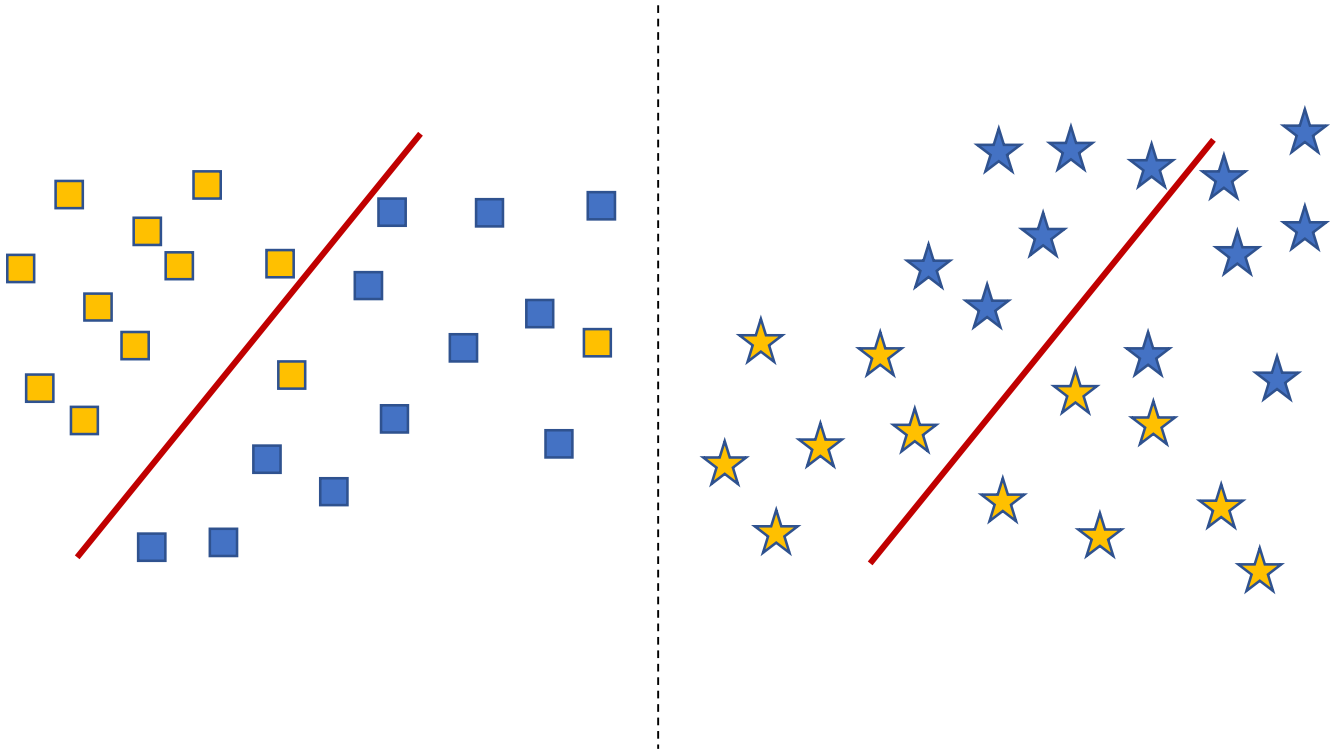
- Occurs when a classifier is trained in one area and deployed in another.



- Example: Bot or Not?
- Squares = Bots in the US
- Stars = Bots in Central Asia

Distributional Shift

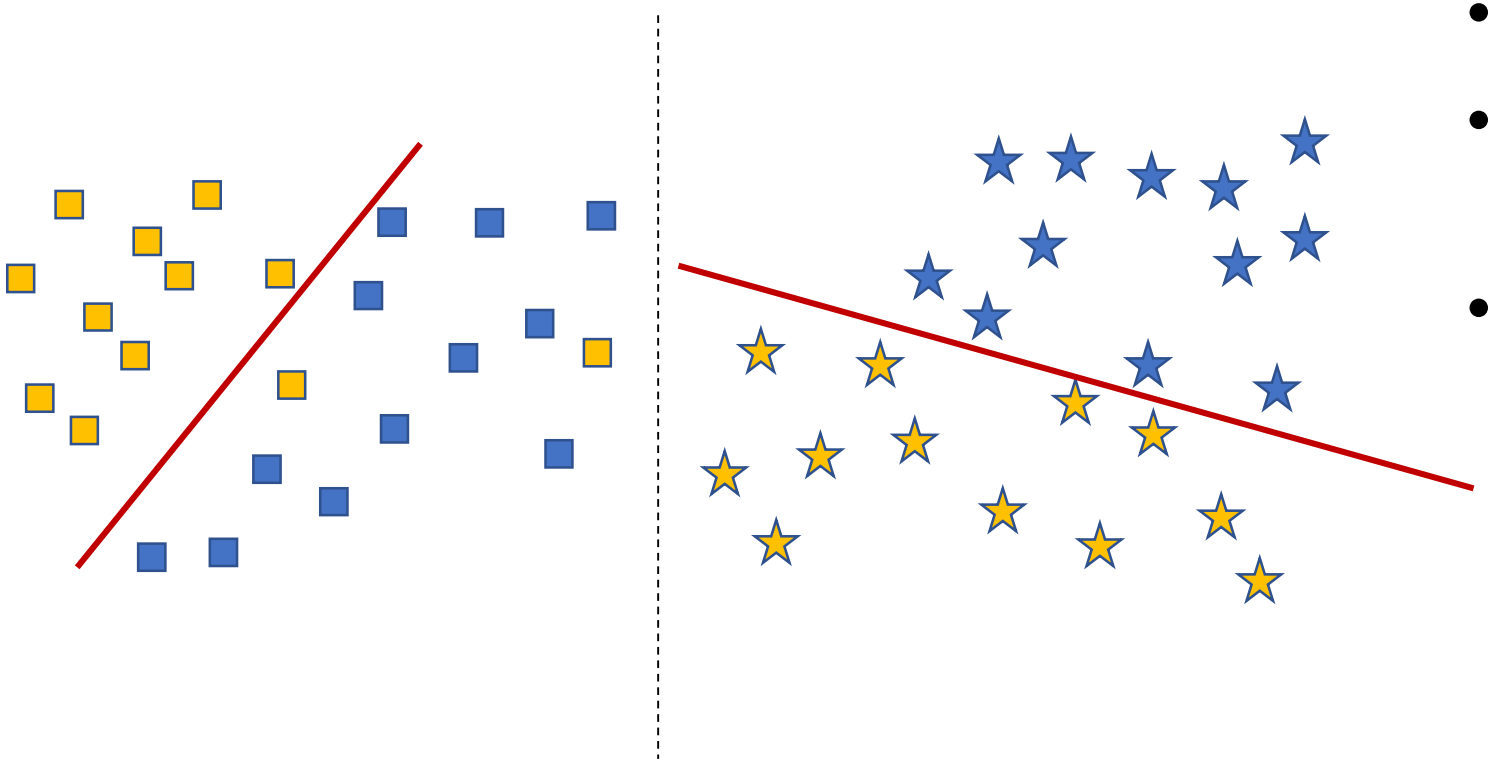
- Occurs when a classifier is trained in one area and deployed in another.



- Example: Bot or Not?
- Squares = Bots in the US
- Stars = Bots in Central Asia

Distributional Shift

- Occurs when a classifier is trained in one area and deployed in another.



- Example: Bot or Not?
- Squares = Bots in the US
- Stars = Bots in Central Asia

Conclusion

- Machine Learning can be a great tool for studying and preventing cybercrime, but is prone to adverse side effects that are often invisible to those deploying them.

Contact Info

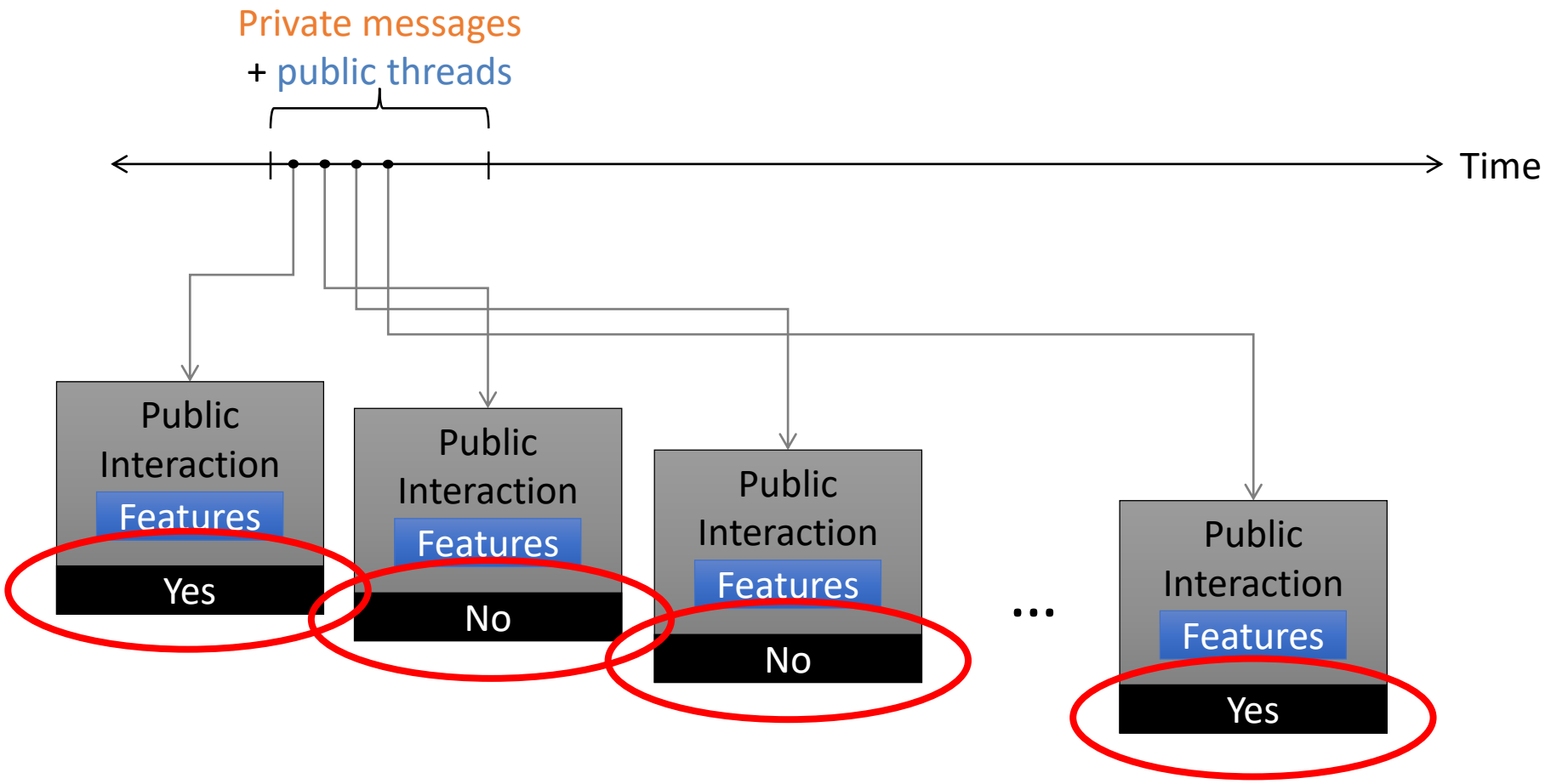
- Rebekah@esat.kueluven.be
- @bekah_Overdorf

- Computer-Supported Cooperative Crime
  - Garg, Afroz, Overdorf, Greenstadt.
- Under the Underground
  - Overdorf, Troncoso, McCoy, Greenstadt
- POTs: Protective Optimization Technologies
  - Overdorf, Balsa, Troncoso, Gurses
- Blogs, Twitter Feeds, and Reddit Comments: Cross-domain Authorship Attribution
  - Overdorf, Greenstadt
- How Unique is Your. onion? An Analysis of the Fingerprintability of Tor Onion Services
  - Overdorf, Juarez, Acar, Greenstadt, Diaz.

Backups

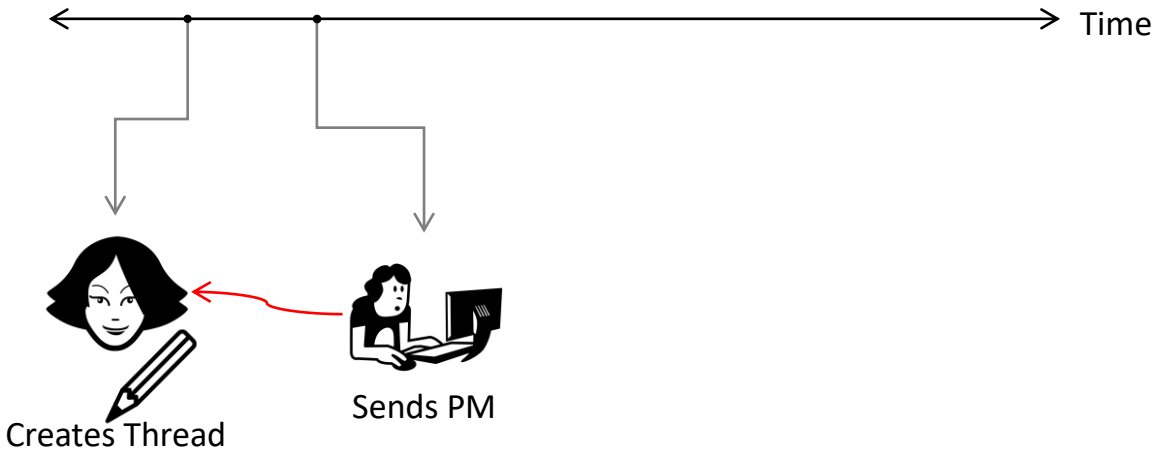
Automated Labeling

Labeling

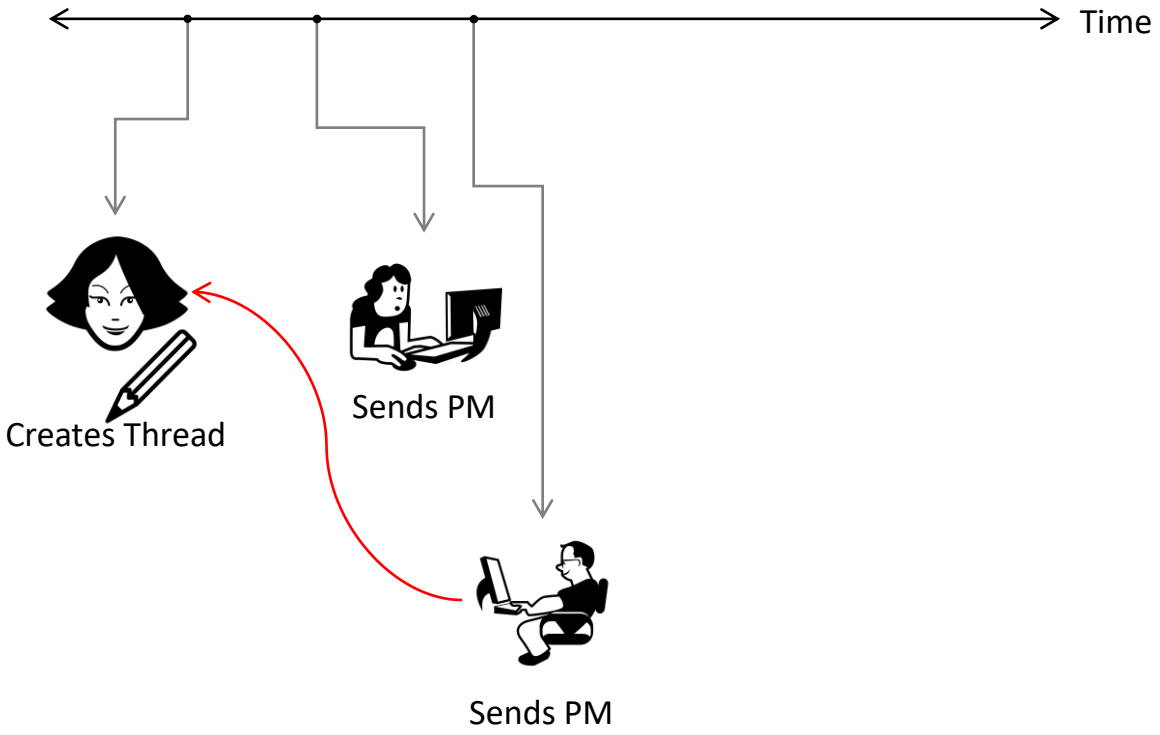

Private messages
+ public threads
Time
Public Interaction
Features
Yes
Public Interaction
Features
No
Public Interaction
Features
No
...
Public Interaction
Features
Yes

Post Example

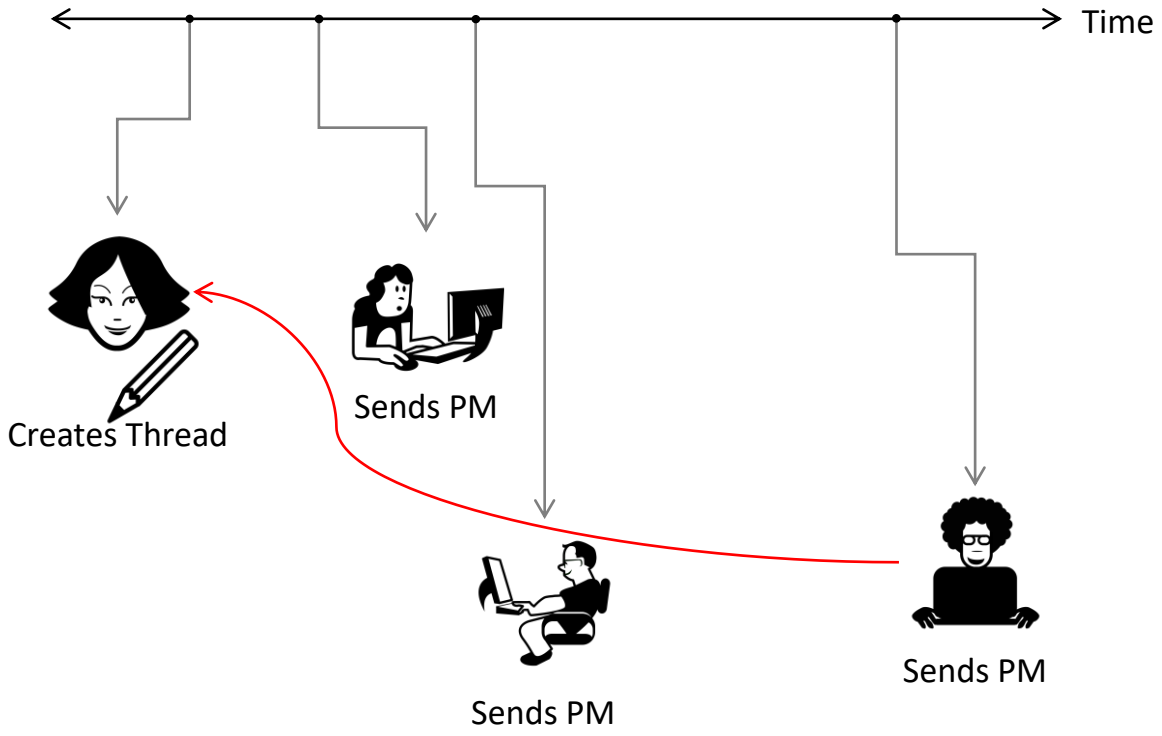
Post Example



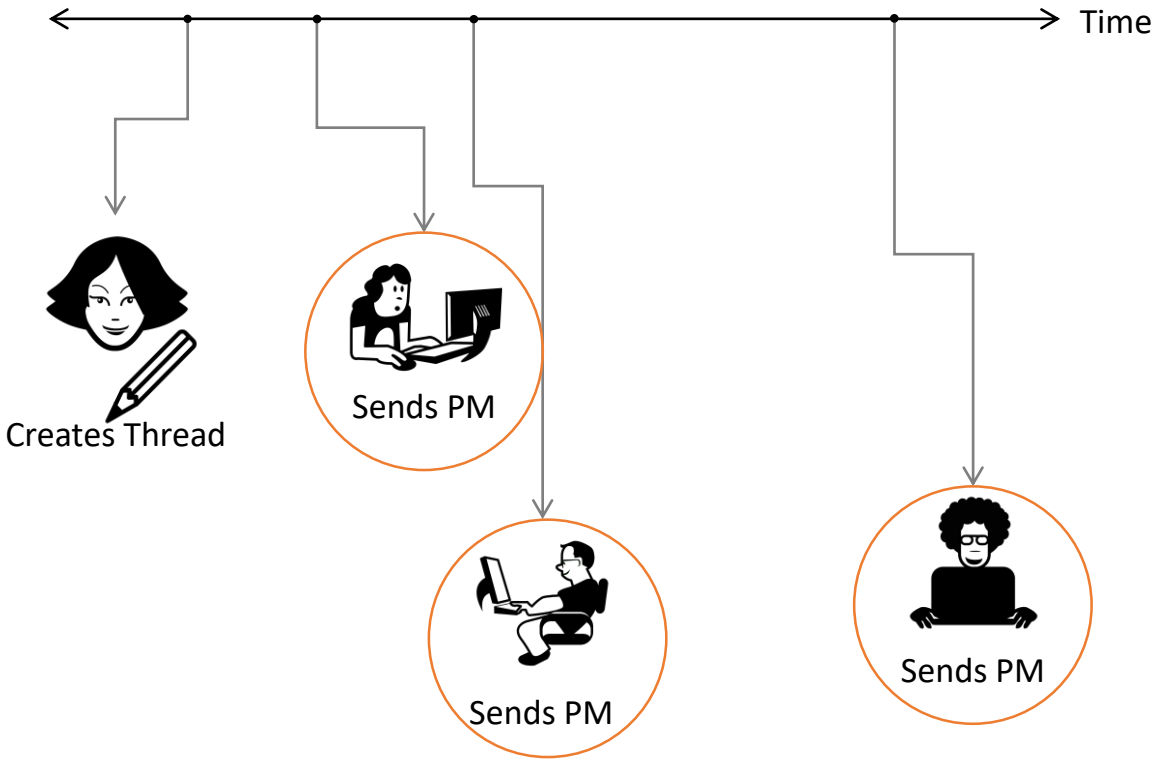
Post Example



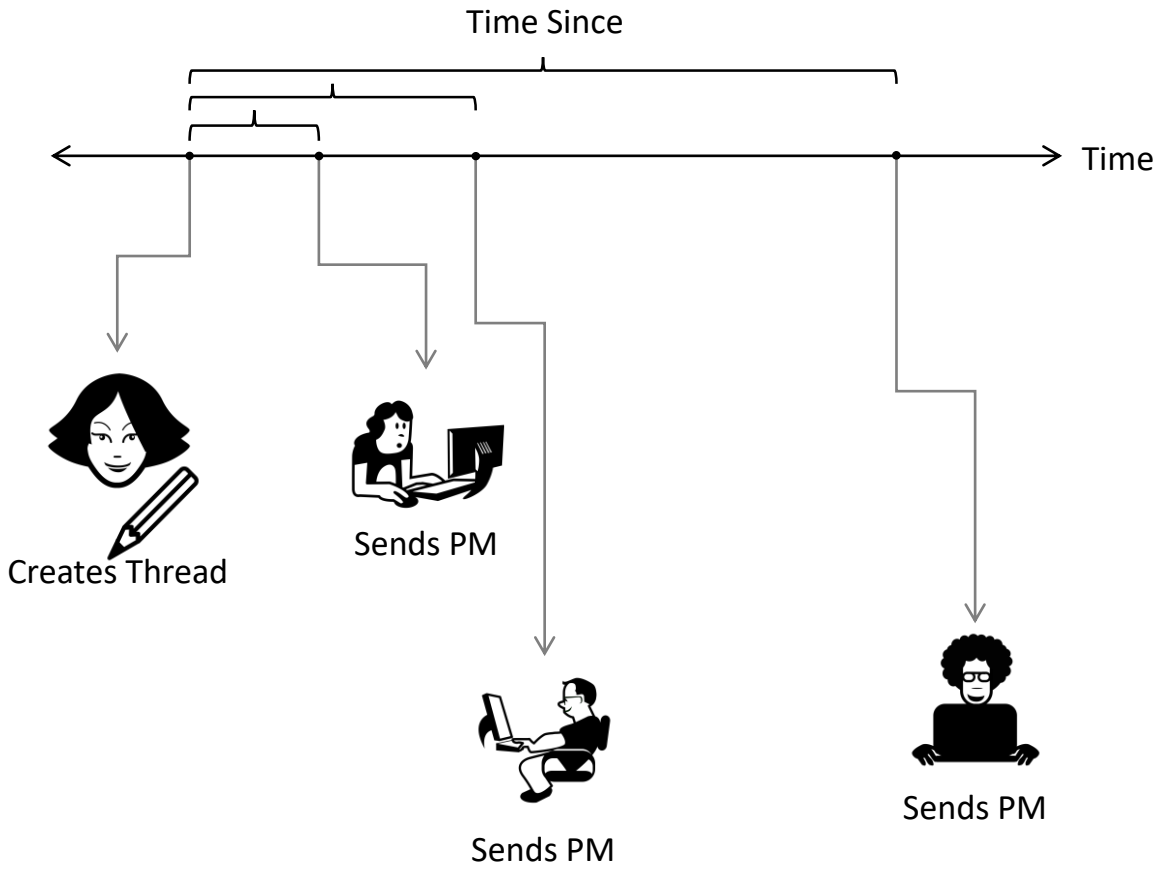
Post Example



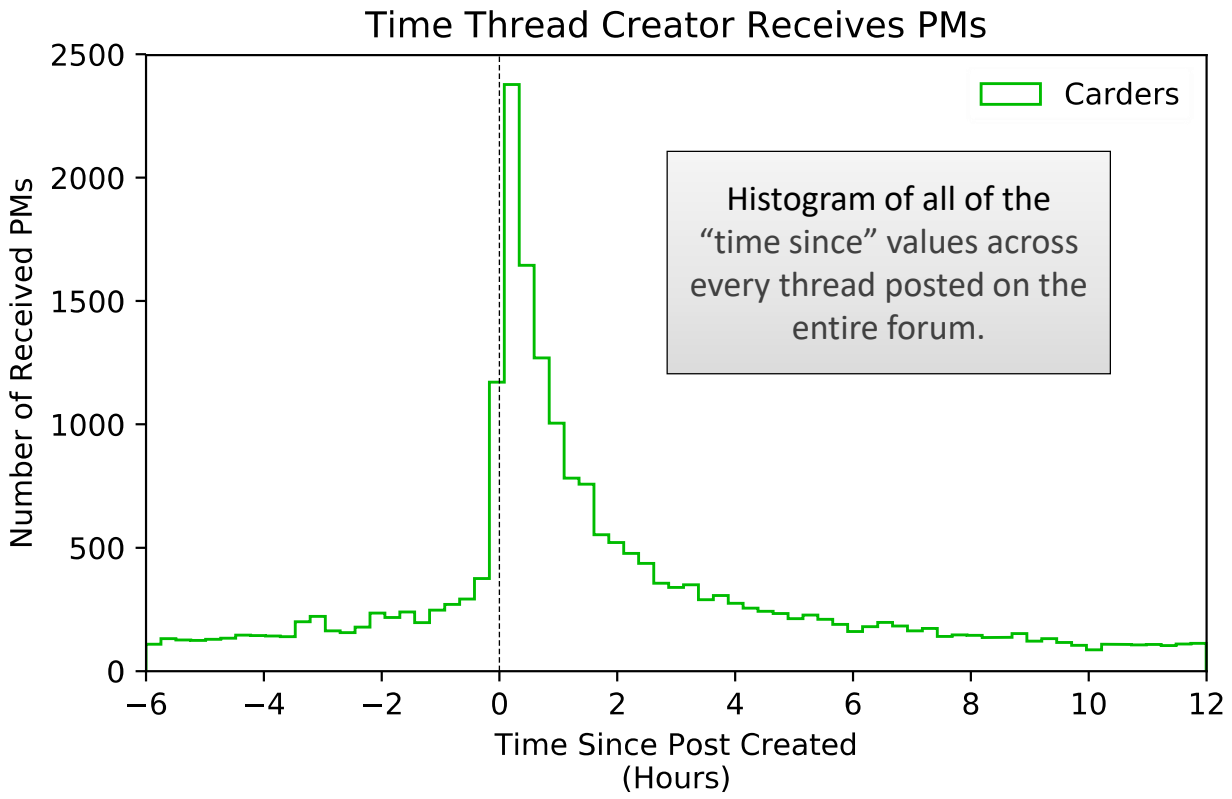
Post Example



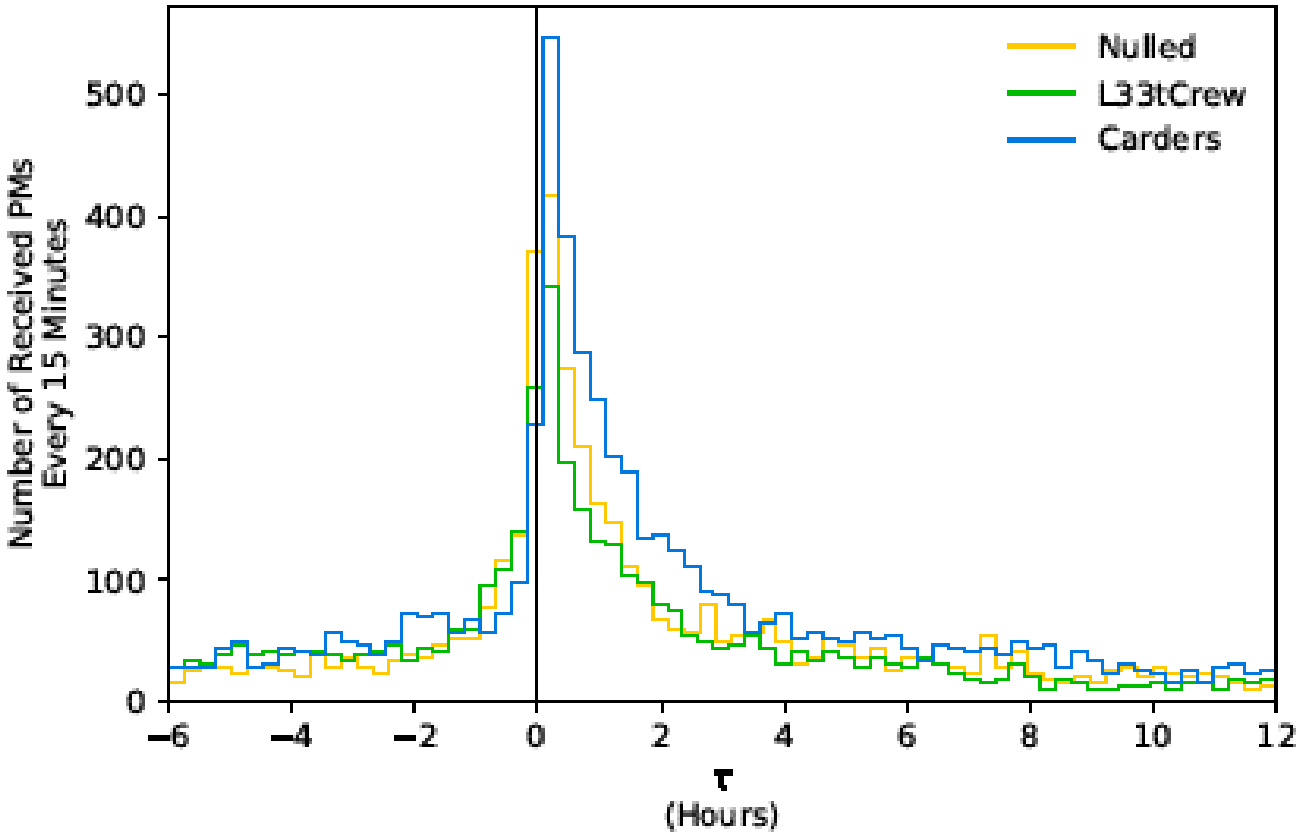
Post Example



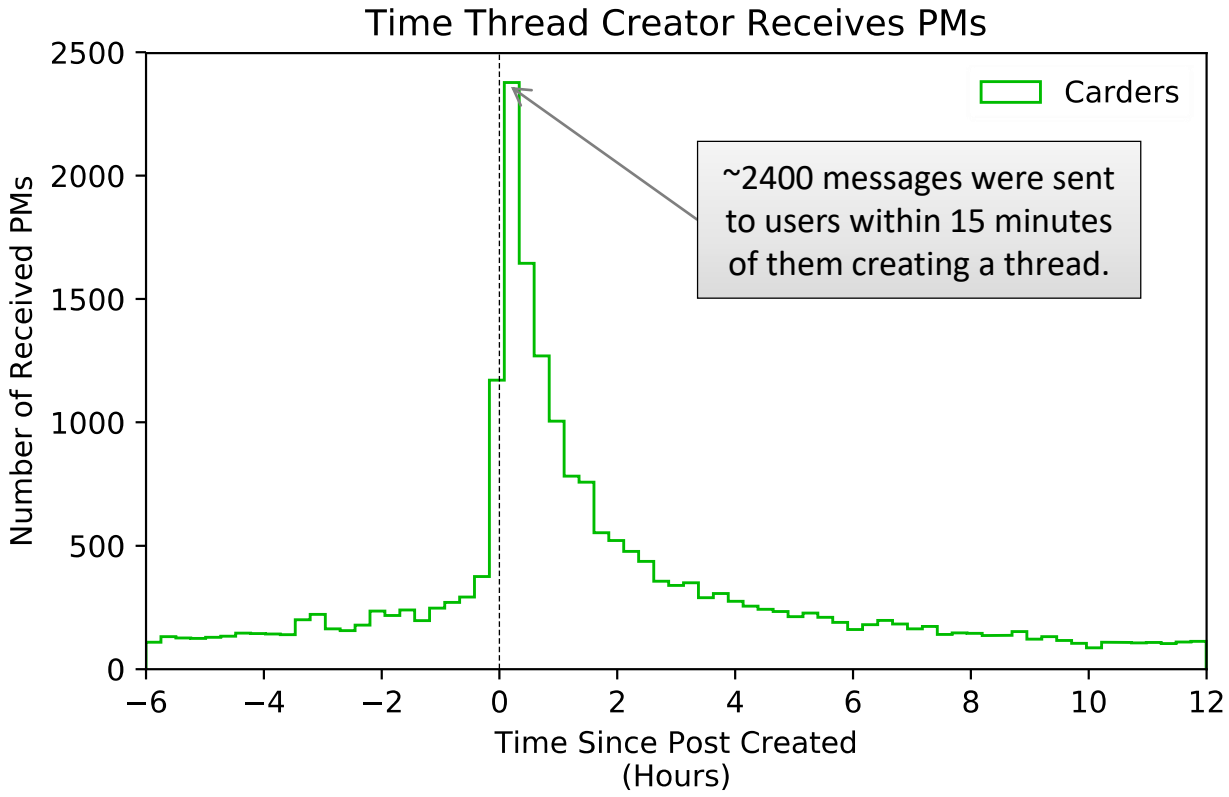
Finding Ground Truth



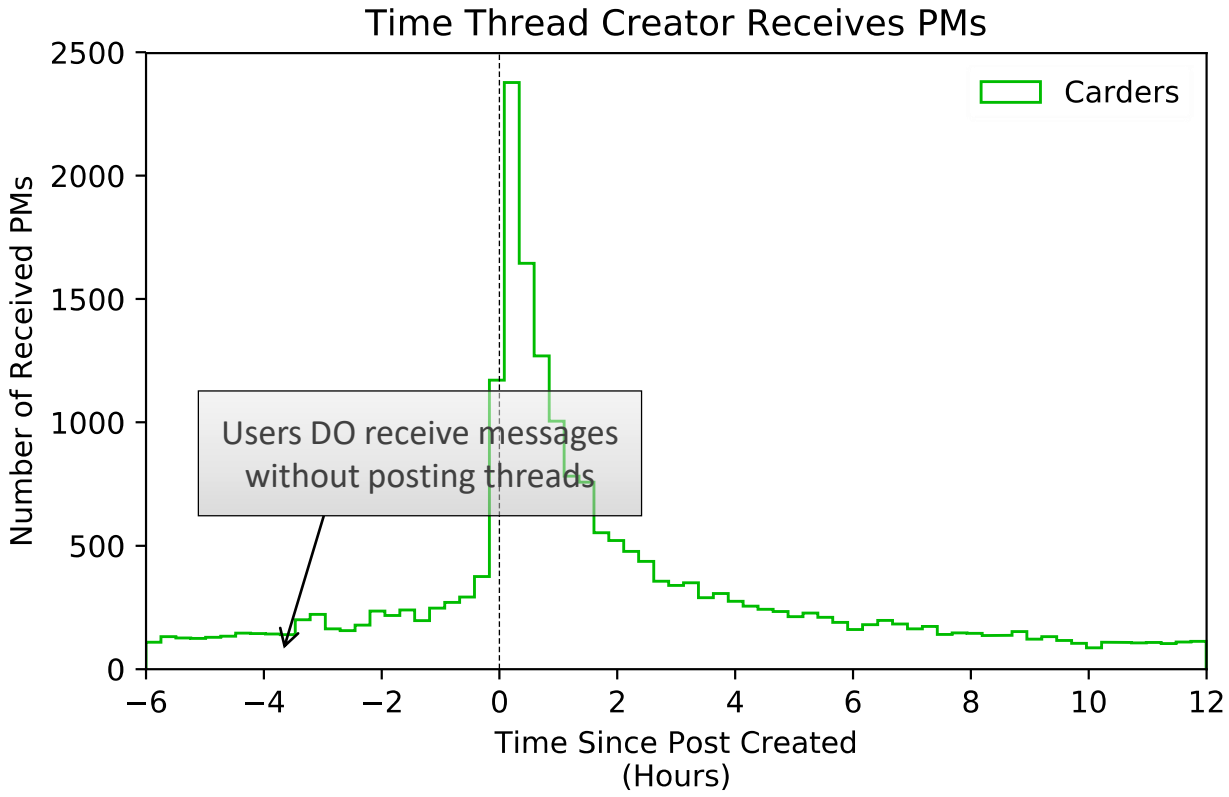
Finding Ground Truth



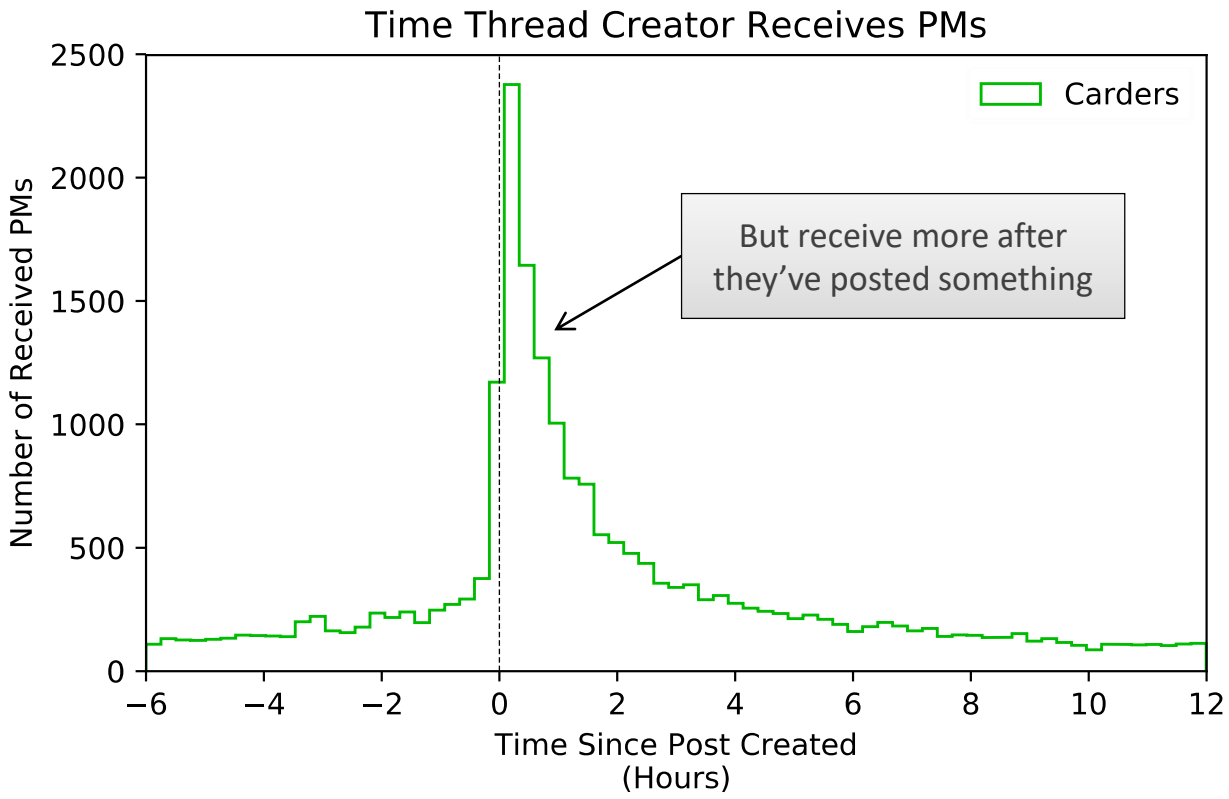
Finding Ground Truth



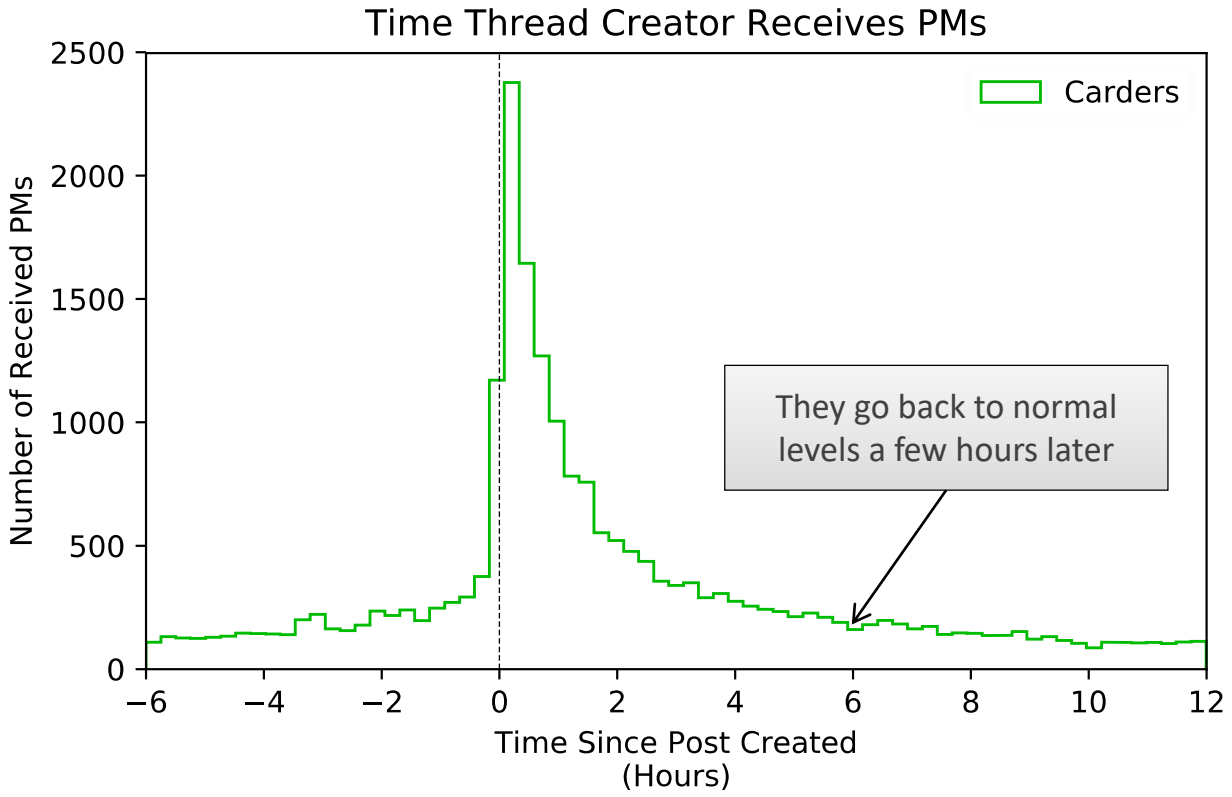
Finding Ground Truth



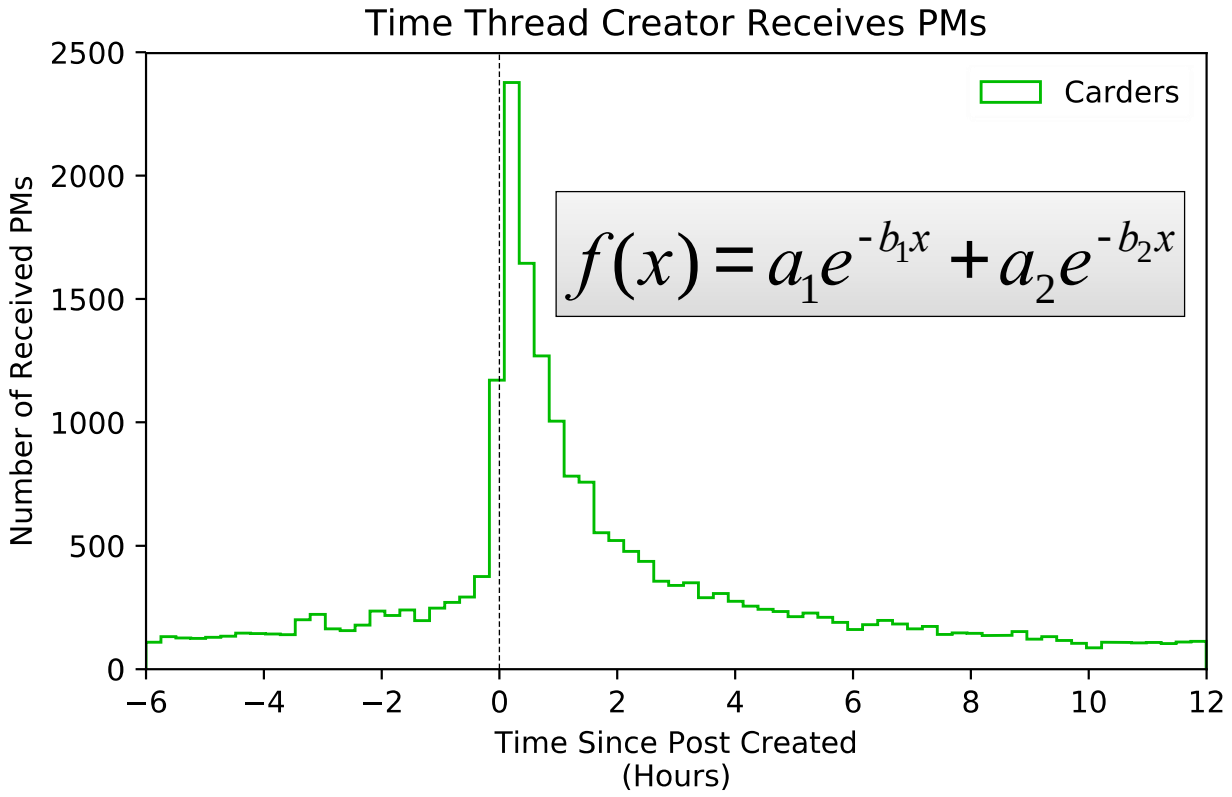
Finding Ground Truth



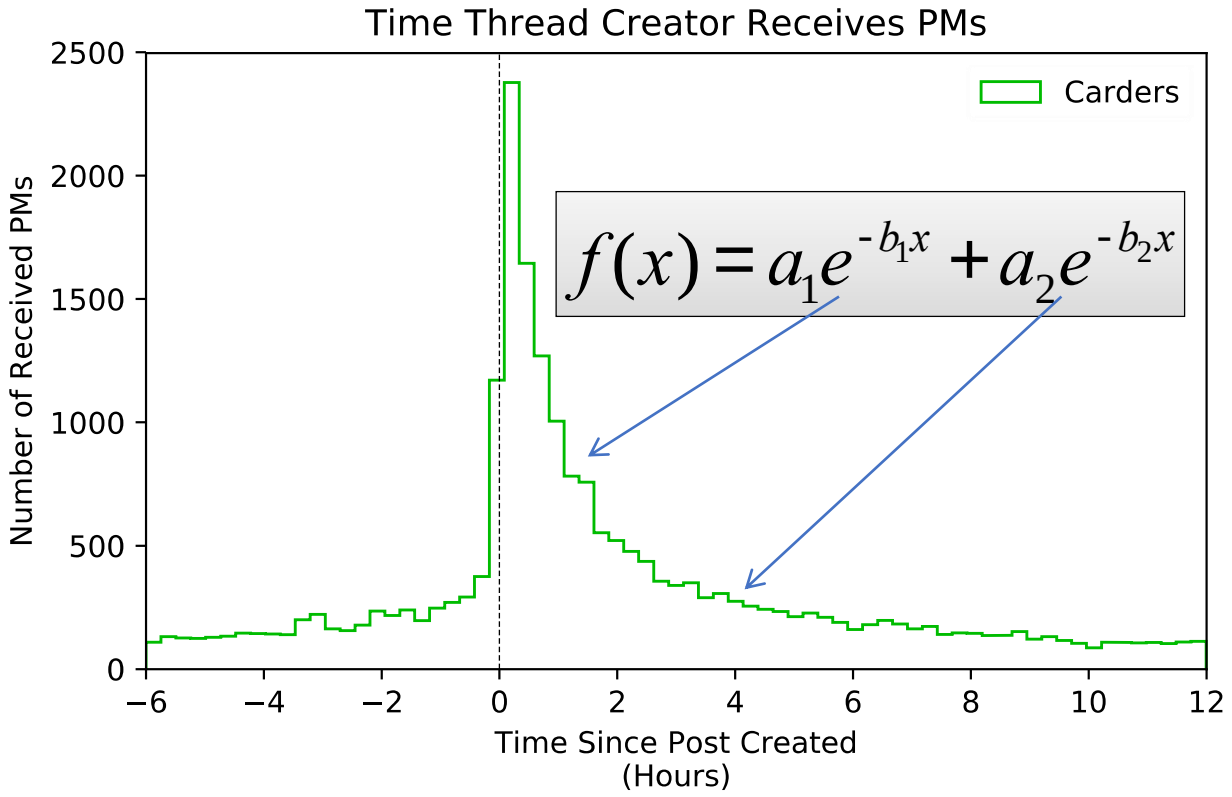
Finding Ground Truth



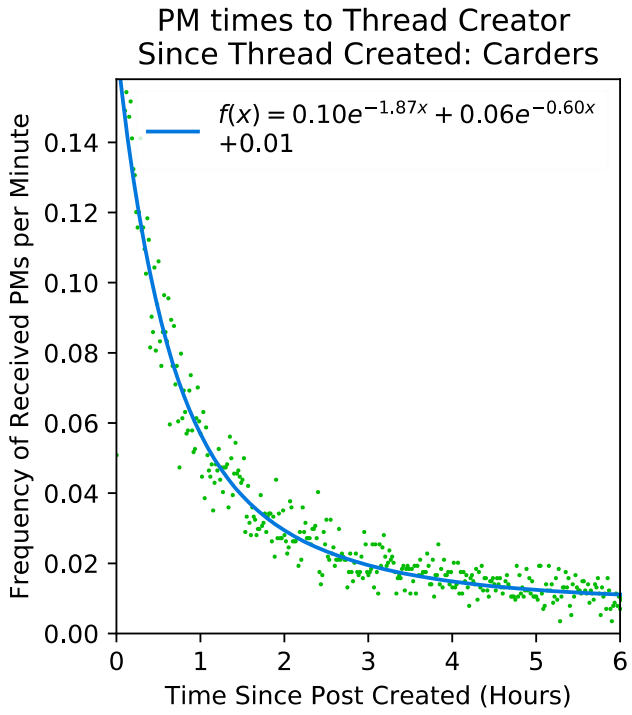
Finding Ground Truth



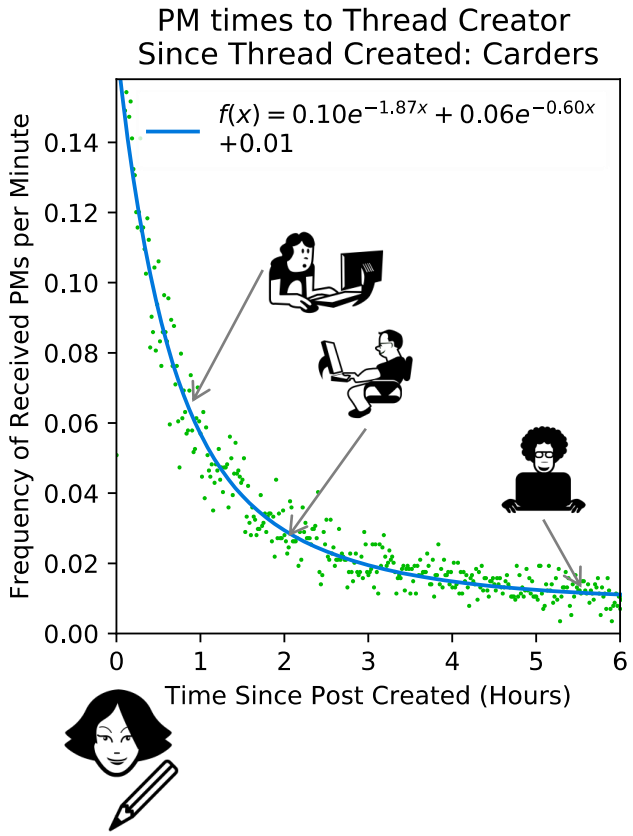
Finding Ground Truth



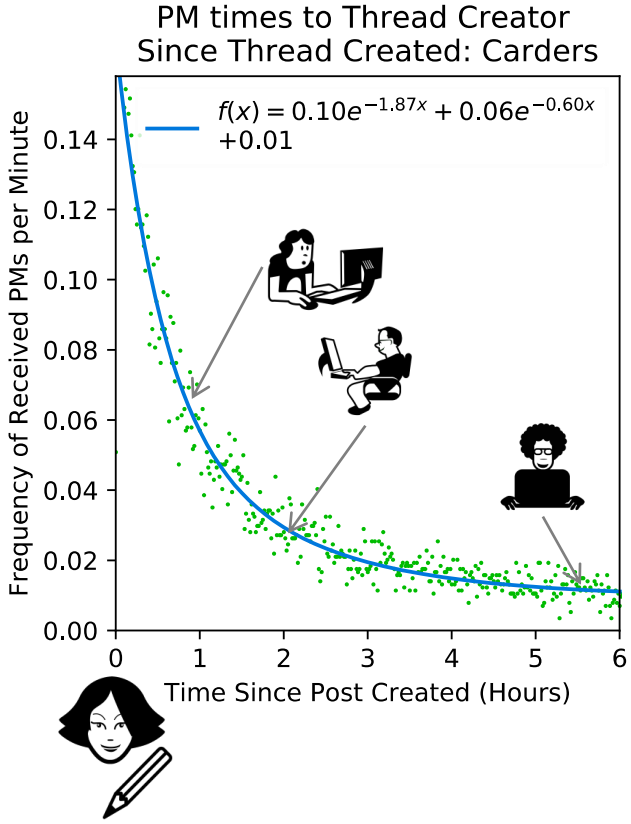
Finding Ground Truth

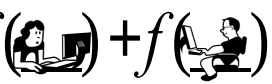


Finding Ground Truth



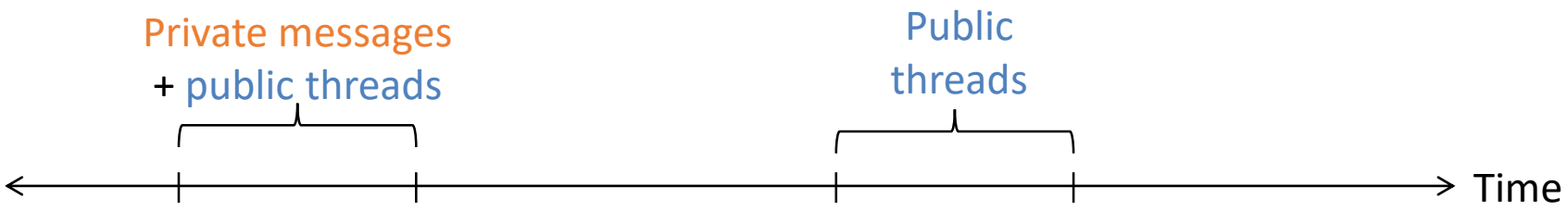
Finding Ground Truth

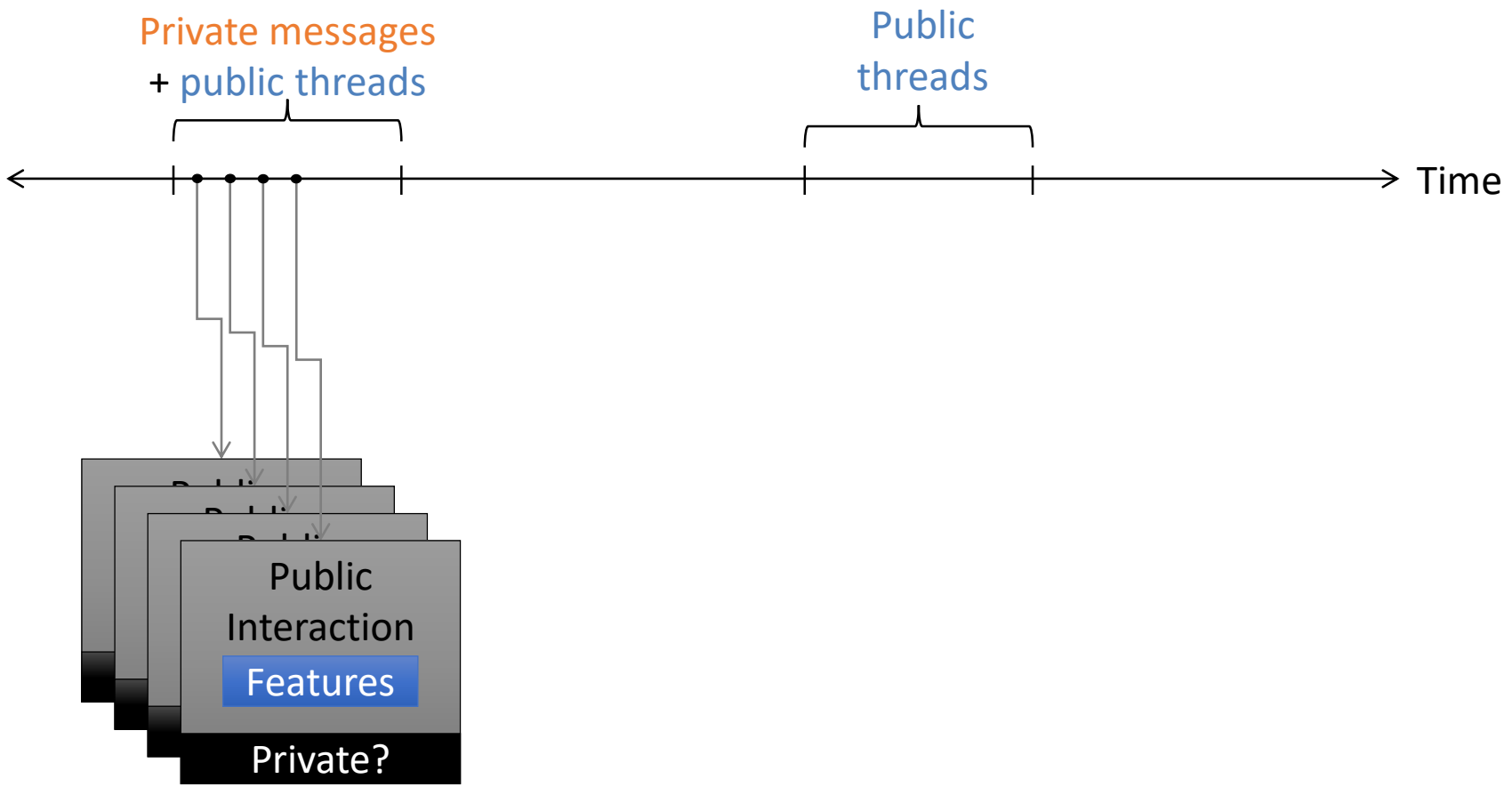


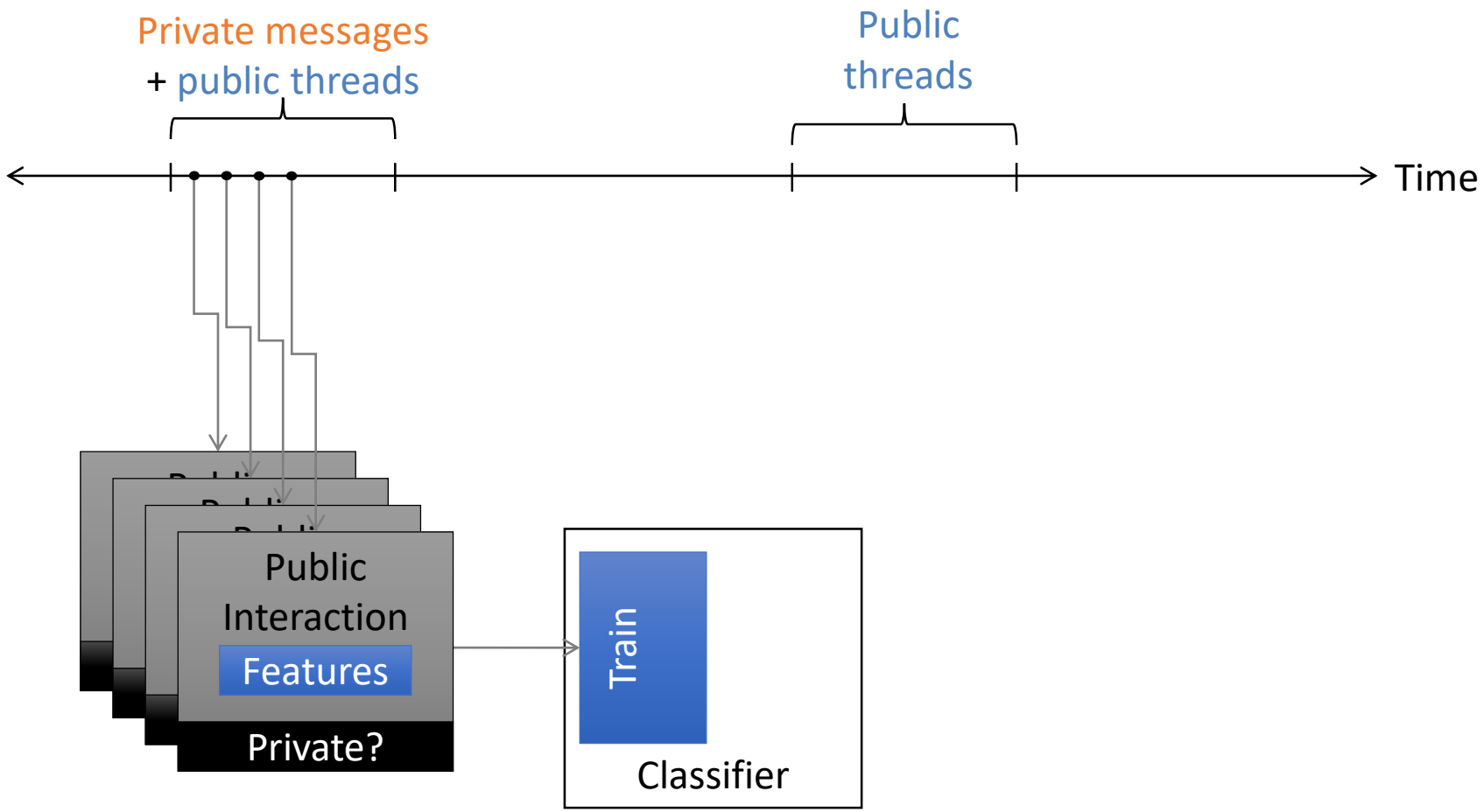
1 if $f($  $)+f($ $)+f($  $) > \theta$

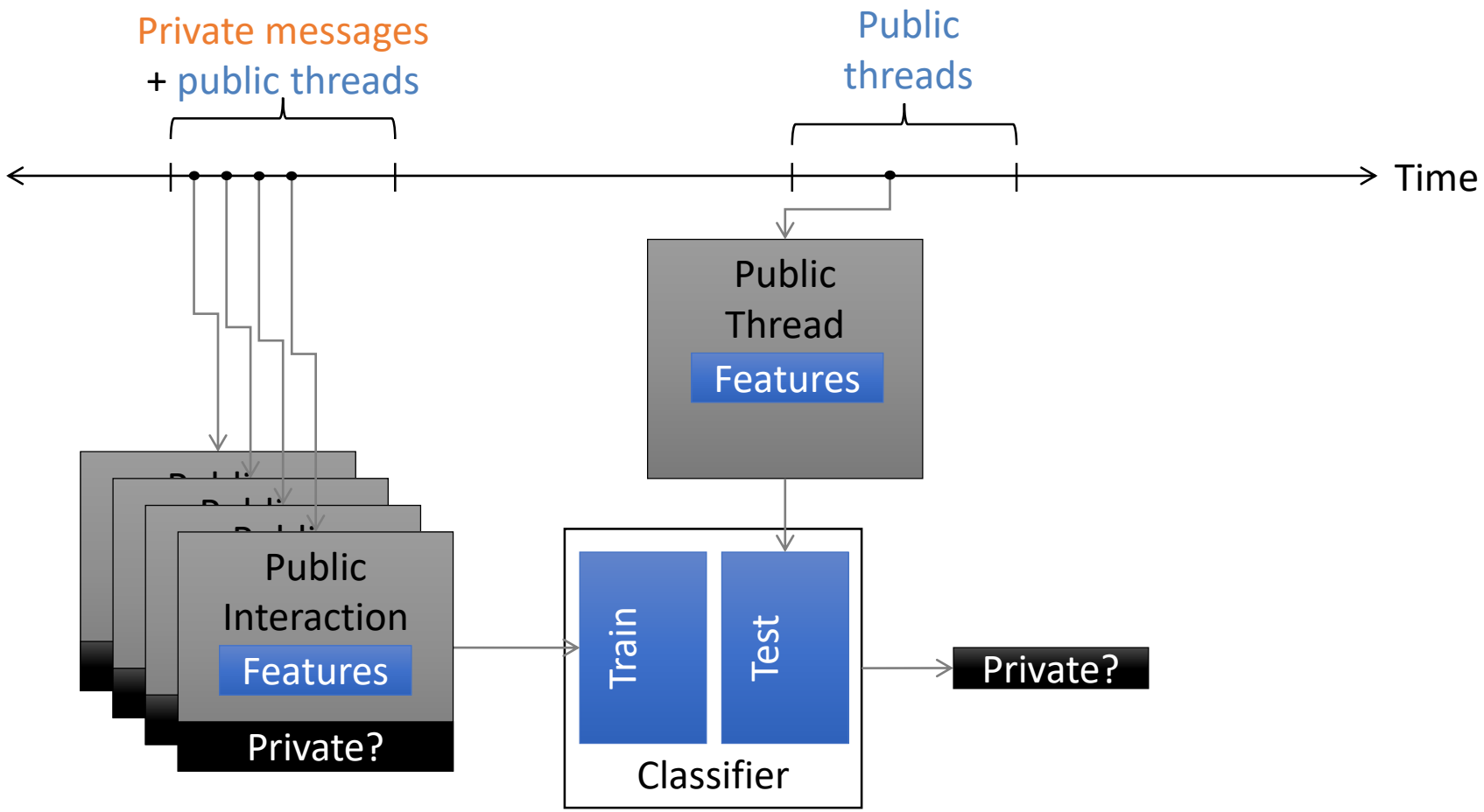
0 otherwise

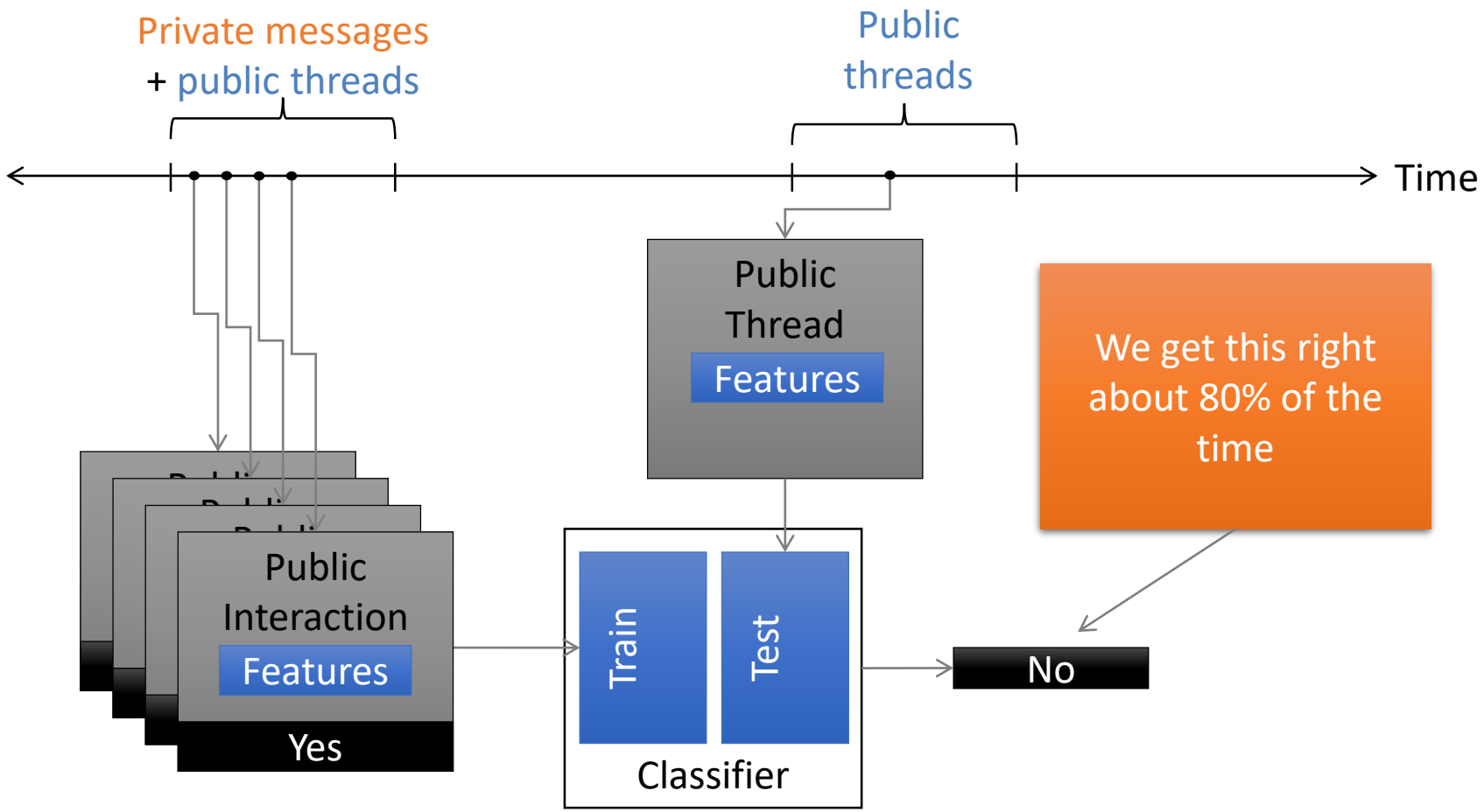

Leak
Time


Private messages
+ public threads
Public
threads
Time


Private messages
+ public threads
Public
threads
Time
Public
Interaction
Features
Private?


Private messages
+ public threads
Public
threads
Time
Public
Interaction
Features
Private?
Train
Classifier


Private messages
+ public threads
Public
threads
Time
Public
Thread
Features
Public
Interaction
Features
Private?
Train
Test
Classifier
Private?

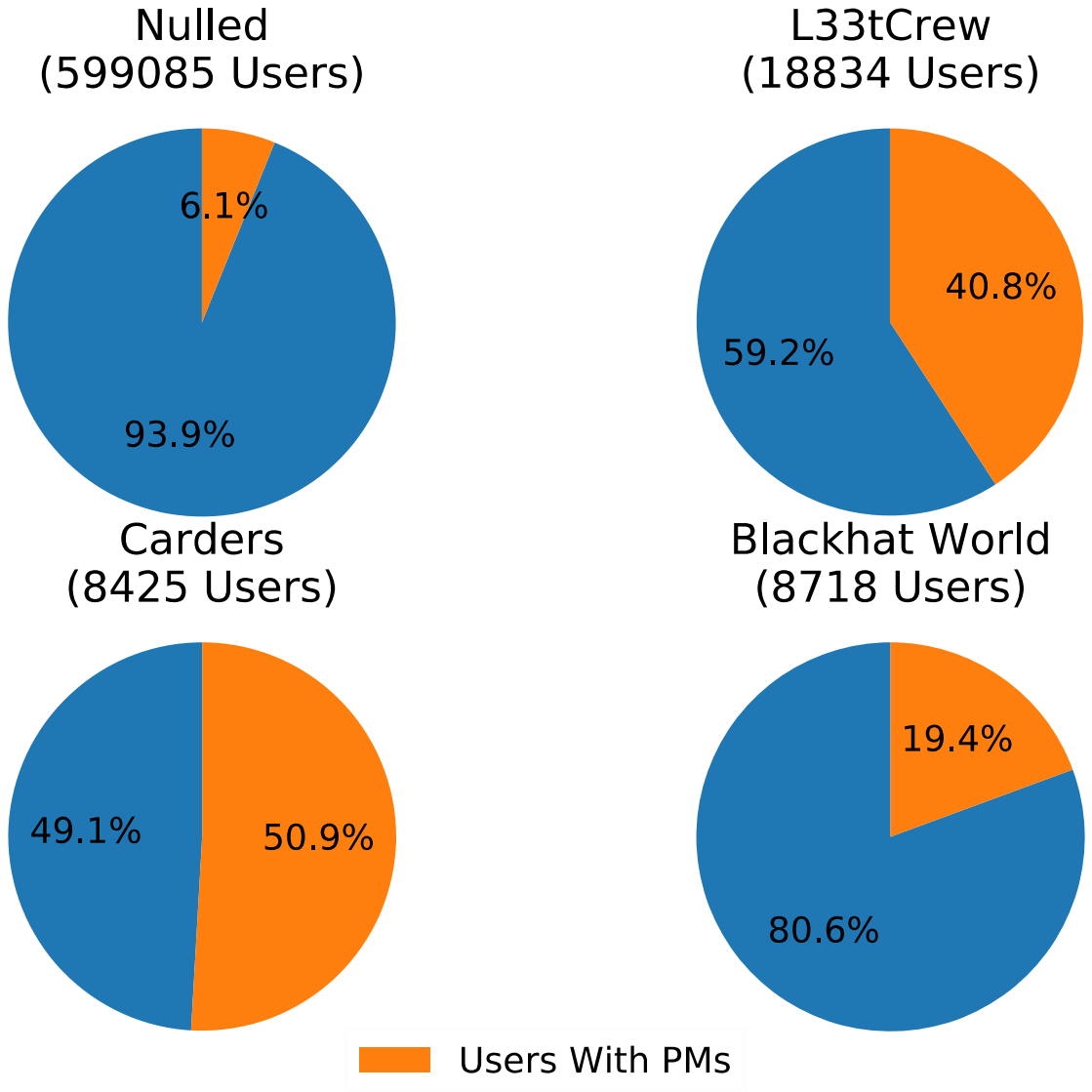

Private messages
+ public threads
Public
threads
Time
Public
Thread
Features
We get this right
about 80% of the
time
Public
Interaction
Features
Yes
Train
Test
Classifier
No

Structure

PM Data Analysis

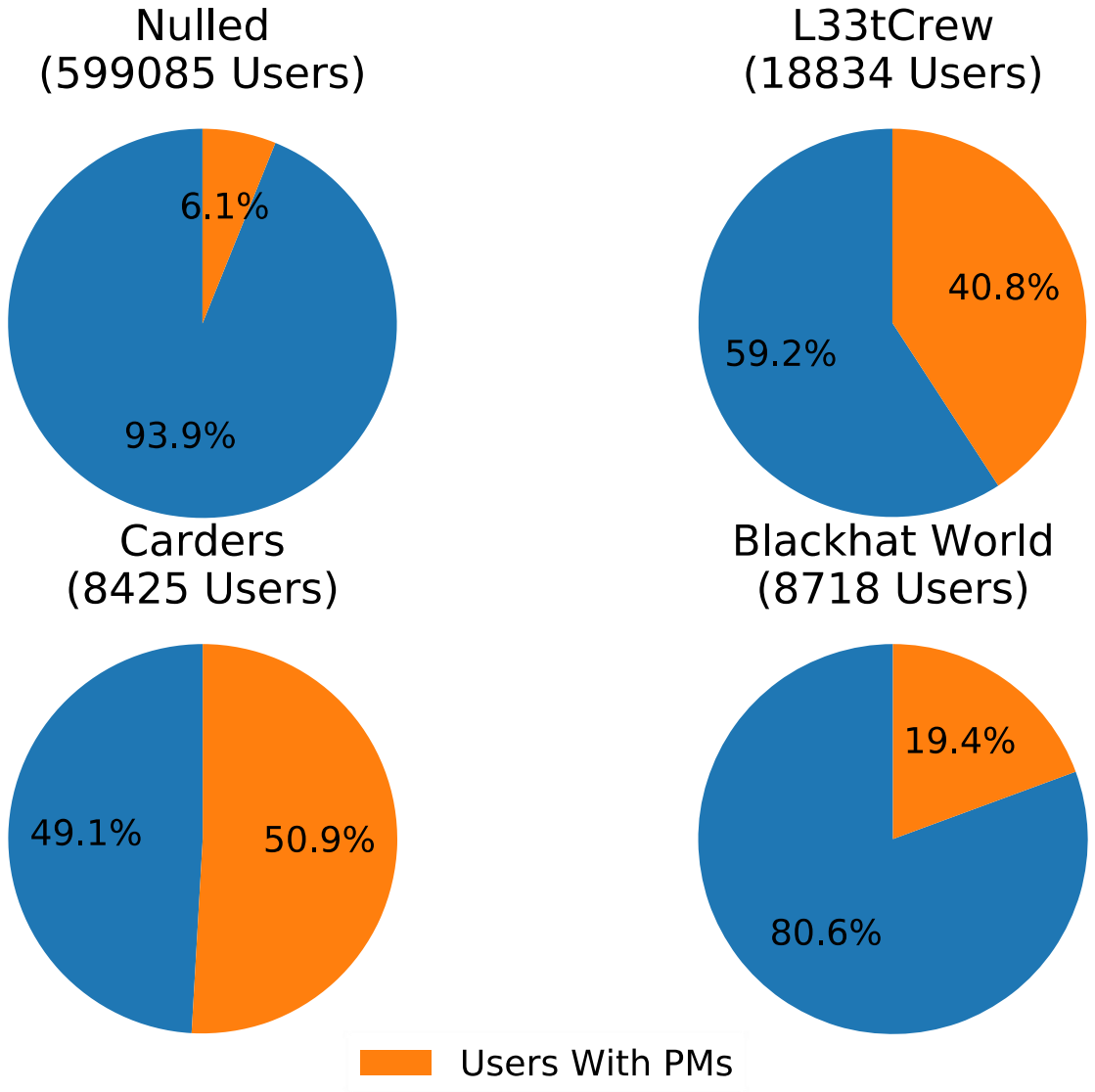
- **Who's important?**
- **Who has the most influence?**
- **What do communities look like?**
- **Is forum-enforced banning effective?**
- **Can we decide which user to remove to be the most disruptive?**
- How does money move?
- How much is a product actually sold for?
- Which items actually sell?
- How does trust flow/scale?
- Can we link accounts?
- …

The Data


Nulled
(599085 Users)
6.1%
93.9%
L33tCrew
(18834 Users)
40.8%
59.2%
Carders
(8425 Users)
49.1%
50.9%
Blackhat World
(8718 Users)
19.4%
80.6%
Users With PMs

The Data

- Carders
  - German
  - Carding
- L33tCrew
  - German
  - Carding
- BlackhatWorld
  - English
  - Young at leak
- Nulled
  - English
  - Huge
  - Varried



Motivation

- How does trust scale?
  - How are the forums organized?
- Who is important?
  - Leaders, central members
- How can we disrupt them?
  - Or how can’t we?

Motivation

- **How does trust scale?**
  - **How are the forums organized?**
- Who is important?
  - Leaders, central members
- How can we disrupt them?
  - Or how can’t we?

Community Detection

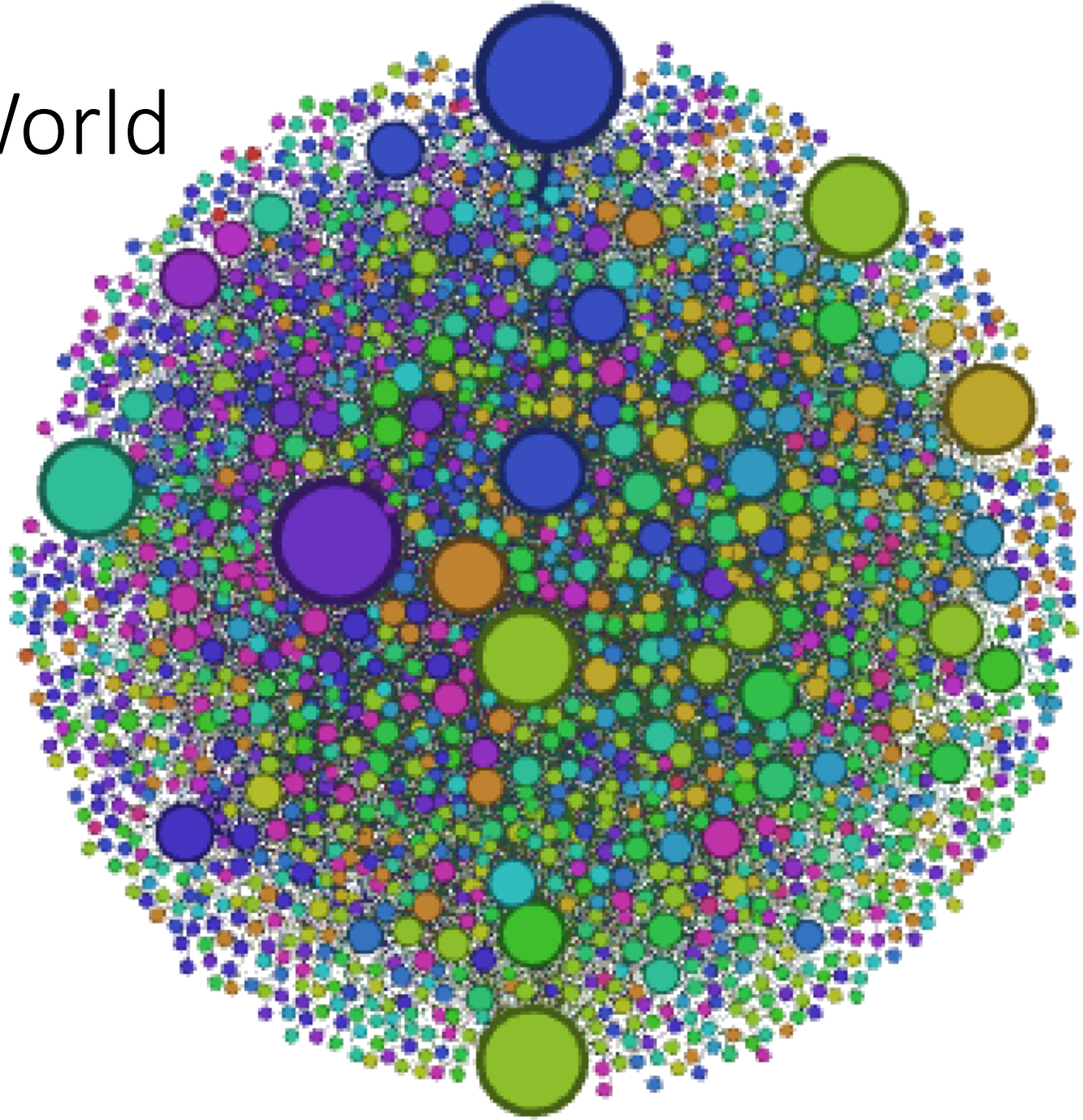
- How are the users organized into communities?
- What do these communities look like?
- What does each specialize in, if at all?

What do these communities look like?

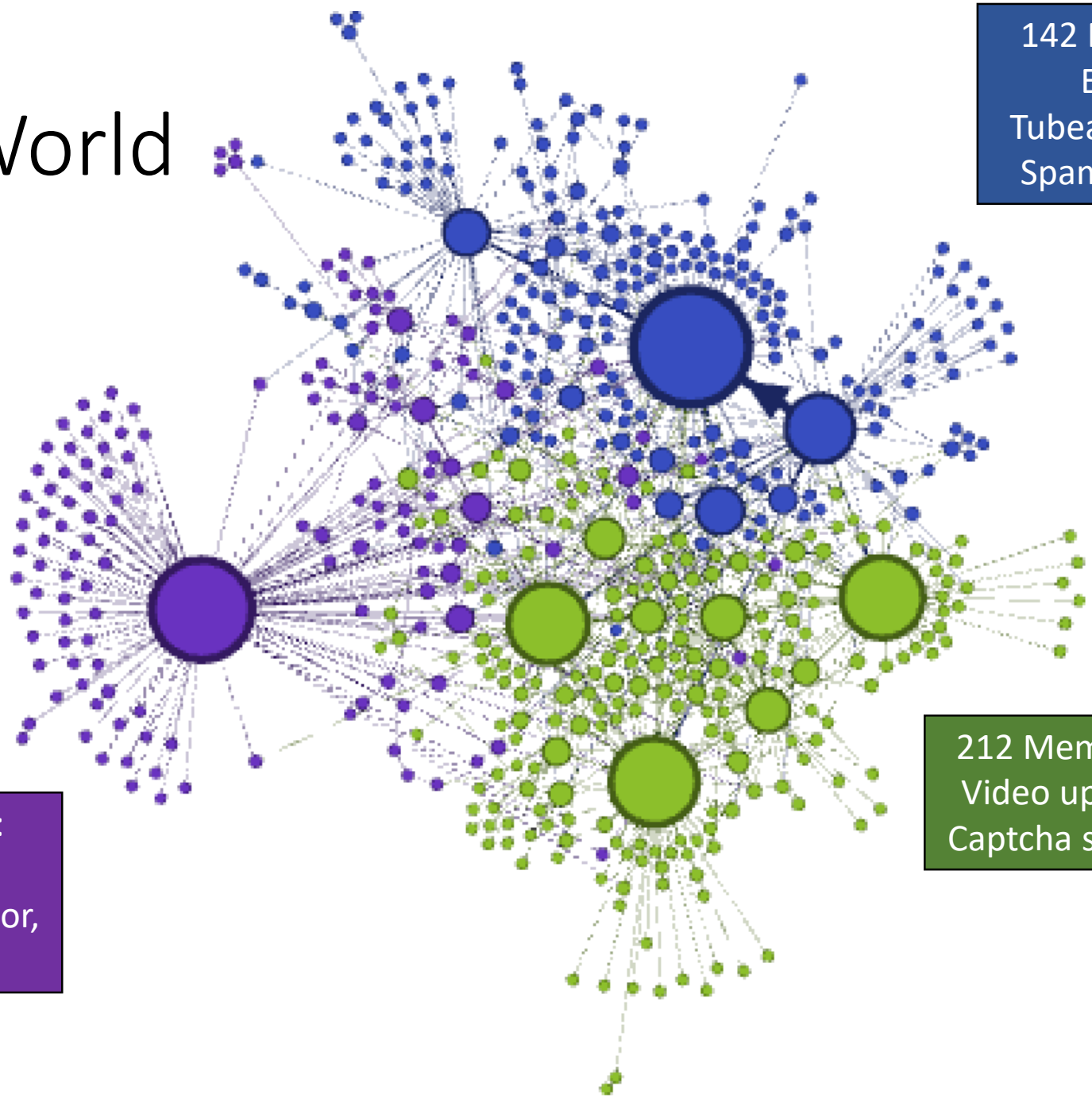
- Dunbar Number (150)
- Structure
  - Mob-like vs Gang-like
- Topics
  - Topics are varied, not uniform, meaning communities specialize.

- Louvain Method for community detection
- LDA for topic modeling

BlackhatWorld



BlackhatWorld

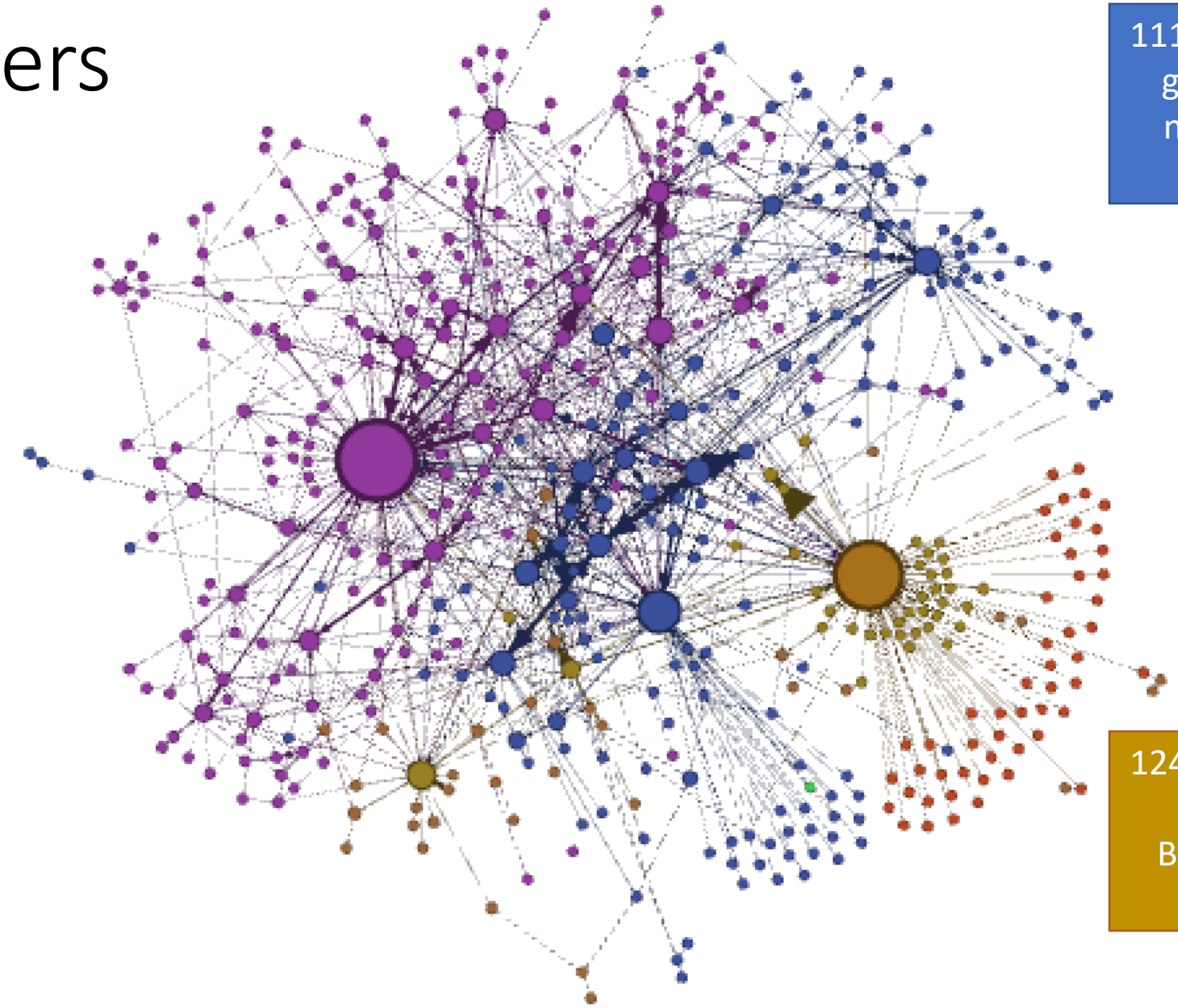


142 Members:
Ebook,
Tubeautomator,
Spammer bots

212 Members:
Video upload,
Captcha solving

203 Members:
Bots
(blogger generator,
promoter)

Carders

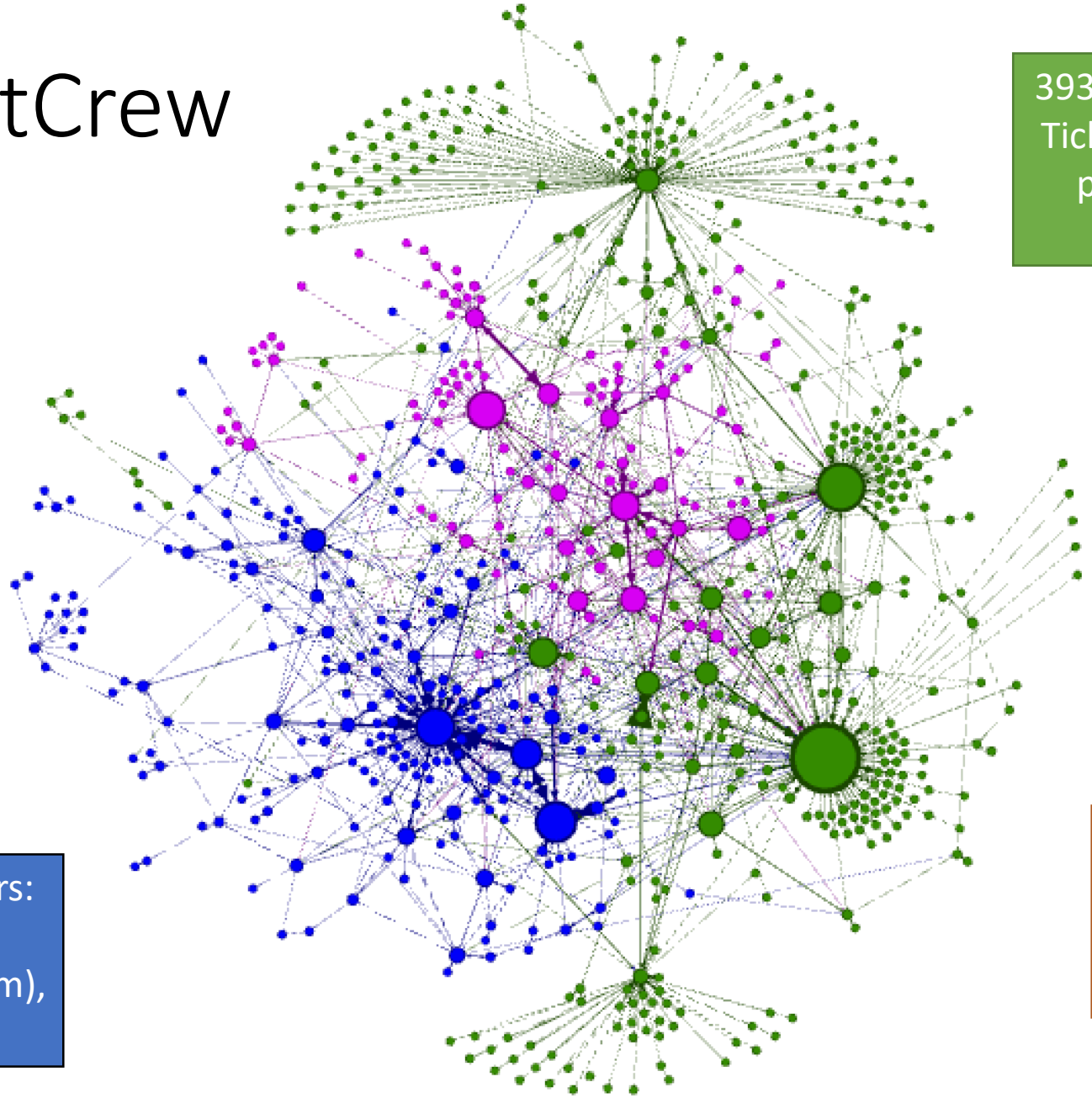


111 Members: gamekeys, notebook, Jabber

197 Members: WebMoney, VPN

124 Members: Jabber, Blackberry, laptop

L33tCrew



393 Members: Tickets, Apple products, Trojans

198 Members: Accounts (paypal, steam), Socks

116 Members: Accounts and keys, Server, Perfume

Motivation

- How does trust scale?
  - How are the forums organized?
- **Who is important?**
  - **Leaders, central members**
- How can we disrupt them?
  - Or how can’t we?

Who's important?

- Degree Centrality
  - Raw number of connections
  - Associated with higher trust
- Betweenness Centrality
  - Number of shortest paths that pass through the node
  - More information
- Closeness Centrality
  - How far is this node from all other nodes
  - Lowest transaction costs
- Eigenvector Centrality
  - How much influence does this node and it's neighbors have

Results

	BlackhatWorld					Carders					L33tCrew				
Cent.	**C**	**B**	**ID**	**OD**	**D**	**C**	**B**	**ID**	**OD**	**D**	**C**	**B**	**ID**	**OD**	**D**
E	0.08	0.66	0.81	0.50	0.71	-0.43	0.79	0.91	0.62	0.77	-0.55	0.85	0.95	0.84	0.91
C		0.33	0.18	0.51	0.37		-0.19	-0.33	-0.11	-0.21		-0.39	-0.51	-0.35	-0.41
B			0.81	0.84	0.88			0.90	0.83	0.90			0.91	0.92	0.94
ID				0.56	0.85				0.71	0.88				0.88	0.96
OD					0.87					0.94					0.96

Results

	BlackhatWorld					Carders					L33tCrew				
Cent.	**C**	**B**	**ID**	**OD**	**D**	**C**	**B**	**ID**	**OD**	**D**	**C**	**B**	**ID**	**OD**	**D**
E	0.08	0.66	0.81	0.50	0.71	-0.43	0.79	0.91	0.62	0.77	-0.55	0.85	0.95	0.84	0.91
C		0.33	0.18	0.51	0.37		-0.19	-0.33	-0.11	-0.21		-0.39	-0.51	-0.35	-0.41
B			0.81	0.84	0.88			0.90	0.83	0.90			0.91	0.92	0.94
ID				0.56	0.85				0.71	0.88				0.88	0.96
OD					0.87					0.94					0.96

- On BlackhatWorld – Everything is correlated.
- On L33tCrew and Carders – Everything but closeness centrality is correlated.
  - Closeness Centrality - How far is this node from all other nodes

Motivation

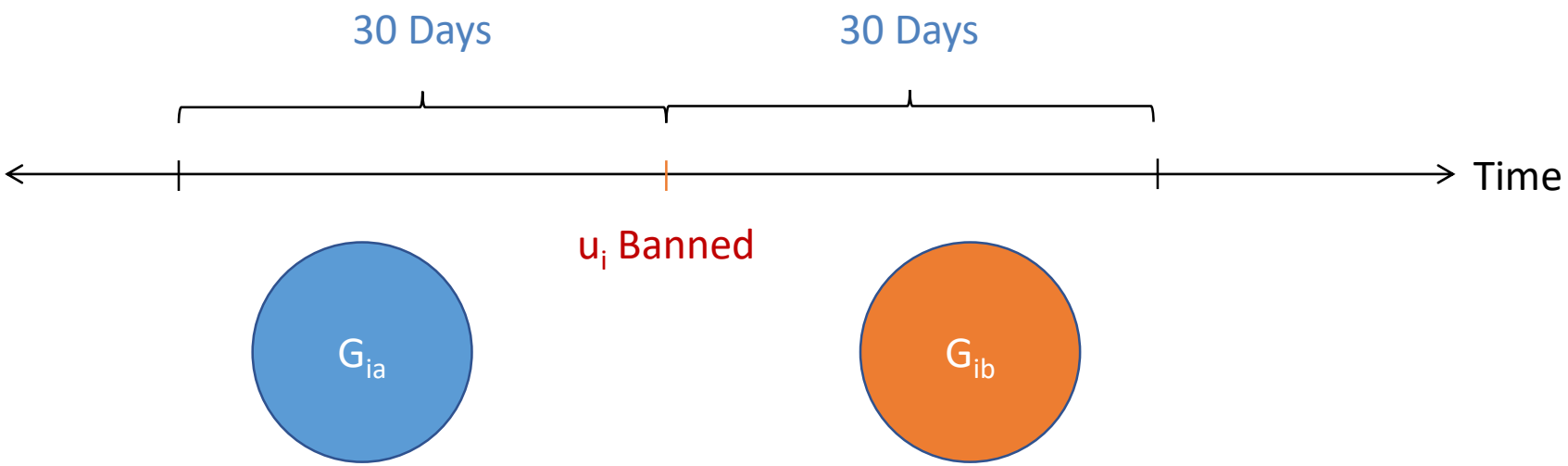
- How does trust scale?
  - How are the forums organized?
- Who is important?
  - Leaders, central members
- **How can we disrupt them?**
  - **Or how can’t we?**

Banning

- Duplicate Accounts
- Ripping
- Spamming

What happens when members are banned?

What happens when members are banned?



"Small World Metrics"
Average Clustering Coefficient
Average Path Length

What happens when members are banned?

Degree Centrality
Betweenness Centrality
Closeness Centrality
Eigenvector Centrality

Results

	BlackhatWorld		Carders		L33tCrew	
CM	**ΔACC**	**ΔAPL**	**ΔACC**	**ΔAPL**	**ΔACC**	**ΔAPL**
Betweenness (B)	-0.39	0.32	-0.12***	-0.05*	-0.05	0.11
Closeness (C)	0.07	-0.12	-0.07**	-0.05*	-0.19*	0.11
Degree (D)	-0.15	0.22	-0.19***	-0.03	-0.06	0.10
Eigenvector (E)	0.07	-0.12	-0.14***	-0.04	-0.01	0.004
p-value: 0.05> * > 0.01 > ** > 0.001 > ***						

Results

	BlackhatWorld		Carders		L33tCrew	
CM	ΔACC	ΔAPL	ΔACC	ΔAPL	ΔACC	ΔAPL
Betweenness (B)	-0.39	0.32	-0.12***	-0.05*	-0.05	0.11
Closeness (C)	-0.07	0.12	-0.07**	-0.05*	-0.19*	0.11
Degree (D)	-0.15	0.22	-0.19***	-0.03	-0.06	0.10
Eigenvector (E)	0.07	-0.12	-0.14***	-0.04	-0.01	0.004
p-value: 0.05> * > 0.01 > ** > 0.001 > ***						

Few Banned Members

Most members banned at the same time

Results

	BlackhatWorld		Carders		L33tCrew	
CM	ΔACC	ΔAPL	ΔACC	ΔAPL	ΔACC	ΔAPL
Betweenness (B)	-0.39	0.32	-0.12***	-0.05*	-0.05	0.11
Closeness (C)	-0.07	0.12	-0.07**	-0.05*	-0.19*	0.11
Degree (D)	-0.15	0.22	-0.19***	-0.03	-0.06	0.10
Eigenvector (E)	0.07	-0.12	-0.14***	-0.04	-0.01	0.004
p-value: 0.05> * > 0.01 > ** > 0.001 > ***						

Few Banned Members

Most members banned at the same time

Results

	BlackhatWorld		Carders		L33tCrew	
CM	ΔACC	ΔAPL	ΔACC	ΔAPL	ΔACC	ΔAPL
Betweenness (B)	-0.39	0.32	-0.12***	-0.05*	-0.05	0.11
Closeness (C)	-0.07	0.12	-0.07**	-0.05*	-0.19*	0.11
Degree (D)	-0.15	0.22	-0.19***	-0.03	-0.06	0.10
Eigenvector (E)	0.07	-0.12	-0.14***	-0.04	-0.01	0.004
p-value: 0.05> * > 0.01 > ** > 0.001 > ***						

Few Banned Members

Most members banned at the same time

- Individuals being banned are not close to other nodes.