

Recherche sur les comportements illicites en ligne via les forums et les réseaux sociaux : méthodes et applications

Professeur Rebekah Overdorf

29 mars 2023

Recherche sur les comportements illicites en ligne via les forums et les réseaux sociaux : méthodes et applications

Professeur Rebekah Overdorf

29 mars 2023

Qu'est-ce qu'un "comportement illicite en ligne" ?

- Activité illégale
- Enfreindre les règles
- Comportement agressif
- Comportement nuisible



Qu'est-ce qu'un "comportement illicite en ligne" ?

- Activité illégale
- Enfreindre les règles
- Comportement agressif
- Comportement nuisible



Vaccine Whistleblower: New vaccine causes sterility in 97% of women!

1 Comment Seen

Réseaux sociaux

Déetecter

Retirer

Prévenir

Forums

Mesure

Perturber

Attribut

Études de cas

Réseaux sociaux

Détection du discours de haine

Détection de compte de bot

Inflation du contenu

Aimés de Facebook

Tendances Twitter

Augmentation du score Reddit

Forums

Forums d'étude

Analyse des structures

Liaison de comptes sur les forums

Étudier les comportements illicites

Cyber guerre

Analyse de contenu



PRIVATE INVESTIGATION SERVICE Dox Anyone Accurate Results Advanced Techniques BTC Accepted

StackFramed

Offline



previously Rigzorra



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 ←



CYBER INVESTIGATIONS

In-depth de-anonymization & investigation

GET STARTED

Forums souterrains



REVEAL THEIR TRUE IDENTITY

Everyone on the web has a unique fingerprint. We take efficient and drastic measures into uncovering the true identity behind

PRIVATE INVESTIGATION SERVICE Dox Anyone Accurate Results Advanced Techniques BTC Accepted



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



CYBER INVESTIGATIONS

In-depth de-anonymization & investigation

GET STARTED

Forums souterrains



REVEAL THEIR TRUE IDENTITY

Everyone on the web has a unique fingerprint. We take efficient and drastic measures into uncovering the true identity behind

PRIVATE INVESTIGATION SERVICE Dox Anyone Accurate R



previously Rigzorra



MEMBER

Posts: 25

Joined: Jun 25, 2017

Reputation: 7

Likes: 6

Credits: 0

Leecher level: 43

HALF YEAR REGISTERED

CYBER INVESTIGATION

In-depth de-anonymization

Forums sout

REVEAL THEIR

Everyone on the web has a unique fingerprint. We take efficient

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

This topic is locked

BTC Accepted

Best way to monetize an email list

👤 Norman_drey · ⏰ Yesterday at 4:02 PM · 📎 email list

Yesterday at 4:02 PM

#1

Norman_drey

Regular Member

INTRO

Joined: May 8, 2022

Messages: 336

Reaction score: 116

What's the best way to monetize an email list, I've been gathering emails from my entertainment blog, so far I've gotten over 100k list, I'm wondering if there could be an easier way to monetize it since I have no interest in paying for bulk email messaging service like malichimp and others

[Advertise on BHW](#)



SimplrHosting

Jr. VIP

Jr. VIP

Joined: Jul 30, 2015

Messages: 175

Reaction score: 59

Website: rb.gy

Omarrr404 said: ⏺

Hi everyone,

I decided to start cold email outreach, and i have a question about gmail warmup as i've read google banned cold email warm up this year.

my gmail is 3 years old and it has been active all this time (sending and receiving a handful of emails each day)
my question is do i still need to warm up my email before sending around 500 emails each day ?

Thank you

No way you can send 500 cold emails each day, this is simply not possible. If you're using paid gsuite account probably around 50 a day, if you're using free @gmail aged account, probably 30 a day.

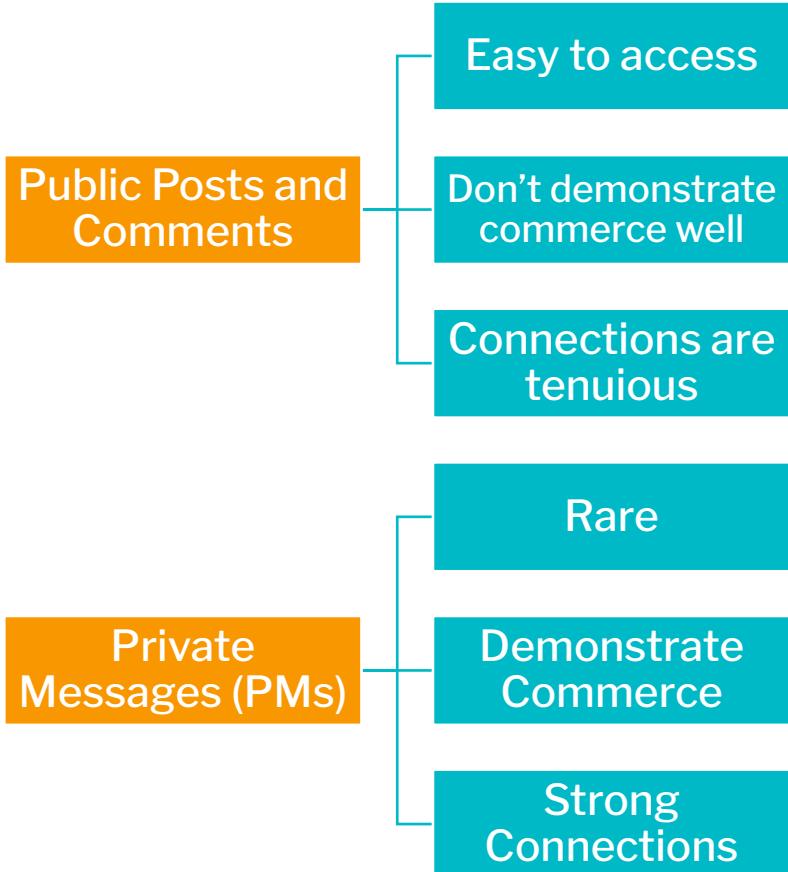
Perfect for Cold Email - 10,000 Targeted B2B Email Leads FROM \$30

⚡ [\$39] AIO-SerpScraper - Harvest Thousands of URLs from Google Super Fast ⚡
My journey to make \$10,000 with affiliates with 0 experience

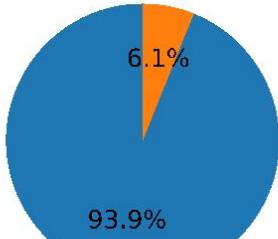
Étudier la structure

- Comment la confiance évolue-t-elle ?
- Comment sont organisés les forums ?
- Qui est important ?
- Dirigeants, membres centraux
- Comment pouvons-nous les perturber ?
- Ou comment ne pouvons-nous pas?

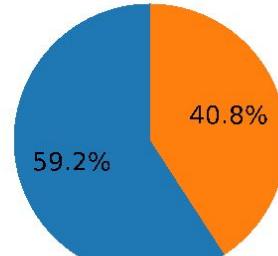
Données



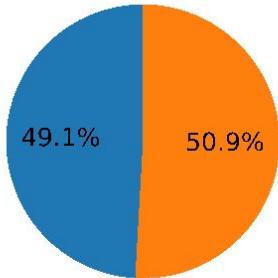
Nulled
(599085 Users)



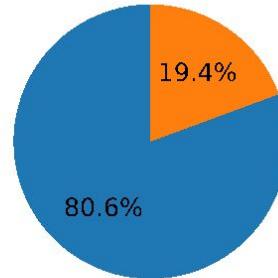
L33tCrew
(18834 Users)



Carders
(8425 Users)



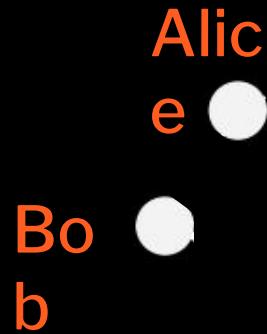
Blackhat World
(8718 Users)



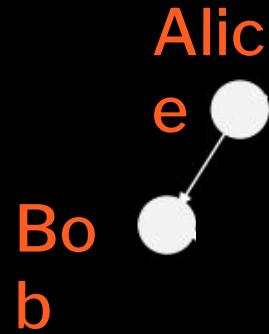
■ Users With PMs

Les données

Construire un graphe social



Construire un graphe social

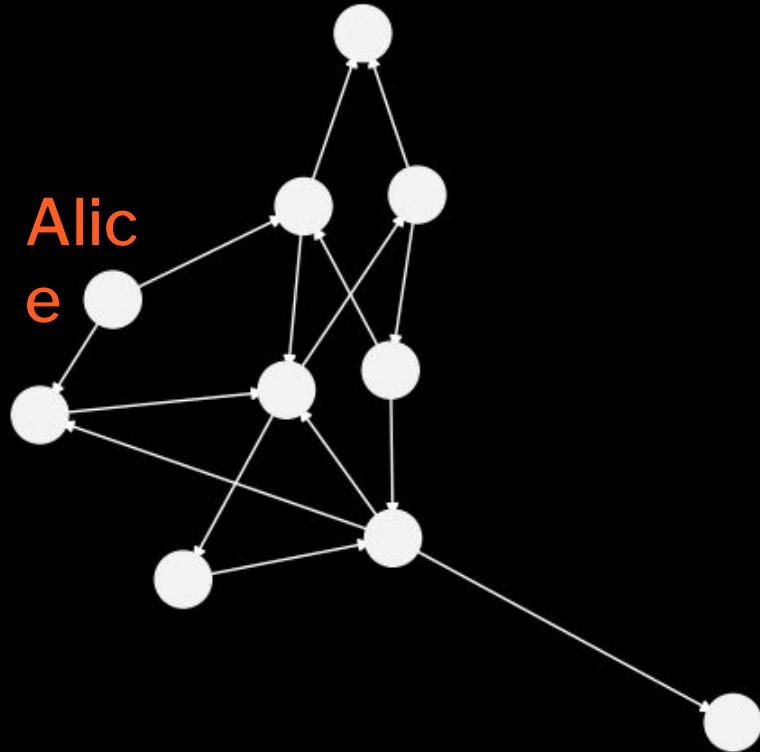


Construire un graphe social

Bo
b

Alic

e



Détection communautaire

Les communautés sont des groupes de nœuds au sein d'un réseau qui sont plus étroitement connectés les uns aux autres qu'aux autres nœuds.

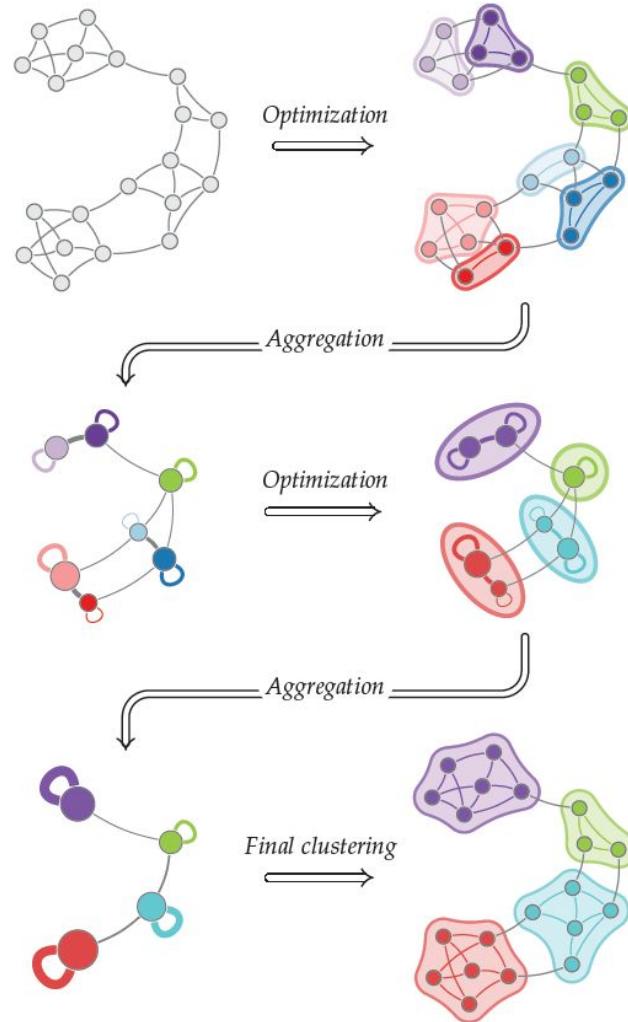
Comment les utilisateurs sont-ils organisés en communautés ?

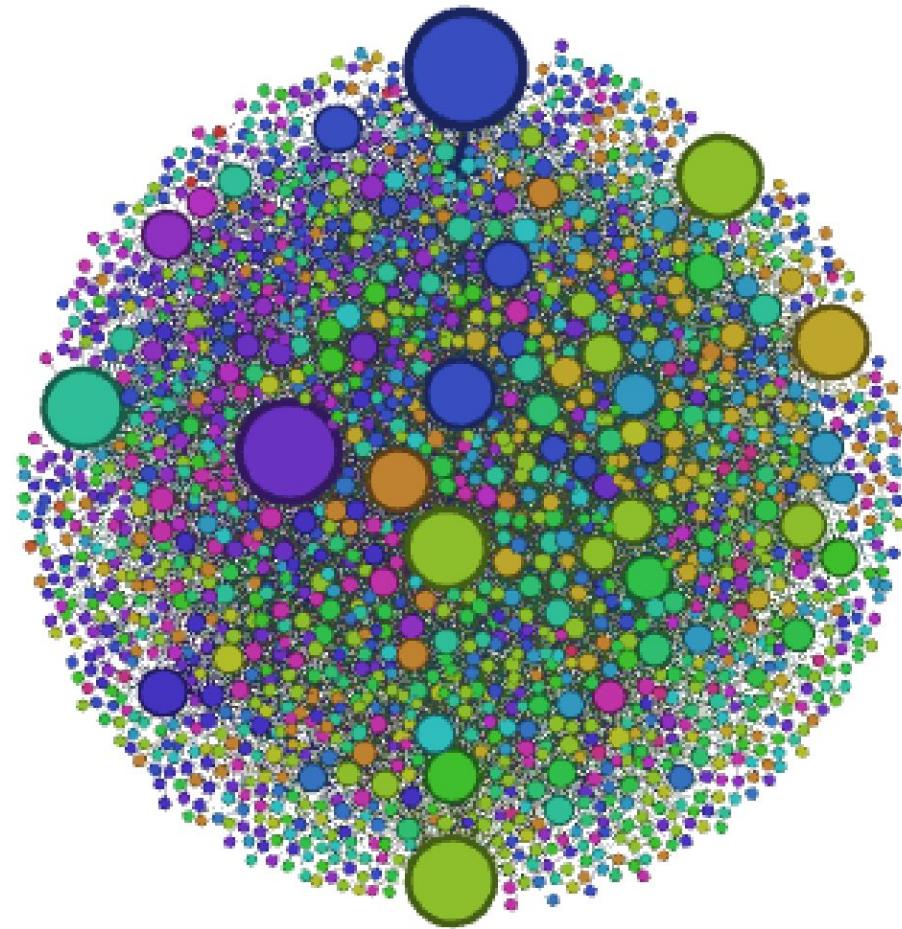
A quoi ressemblent ces communautés ?

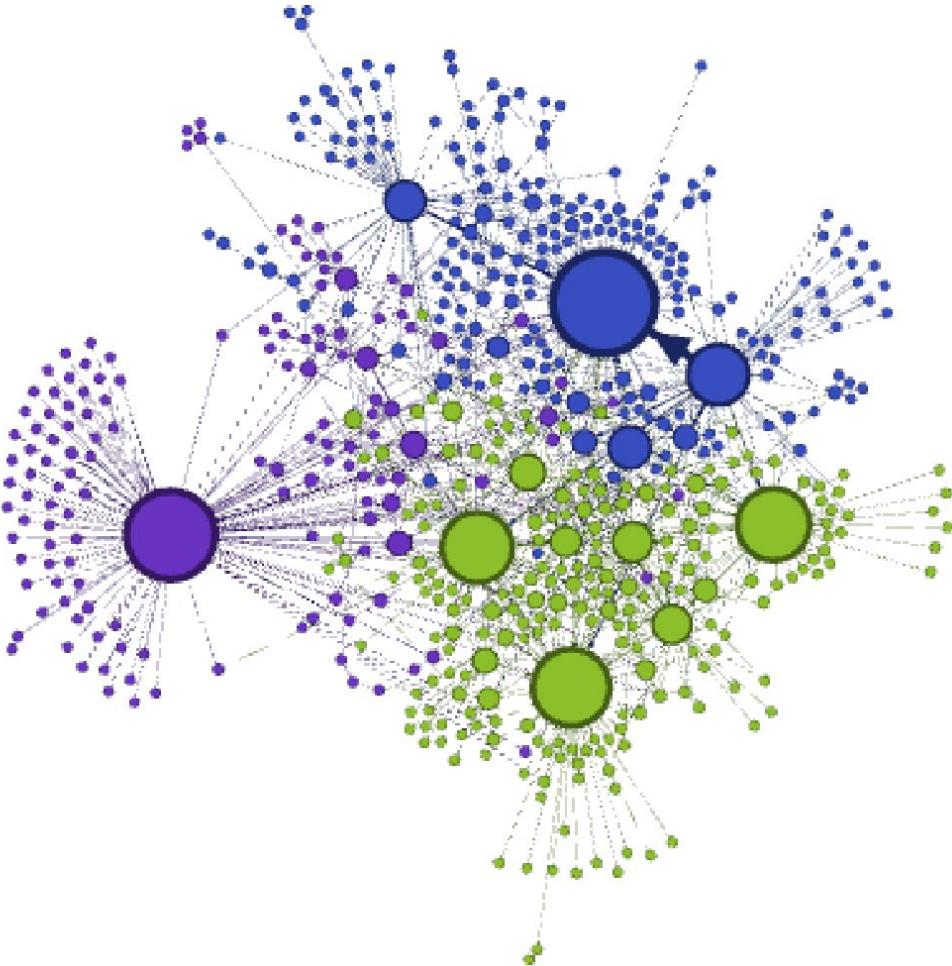
Dans quoi chacun se spécialise-t-il, le cas échéant ?

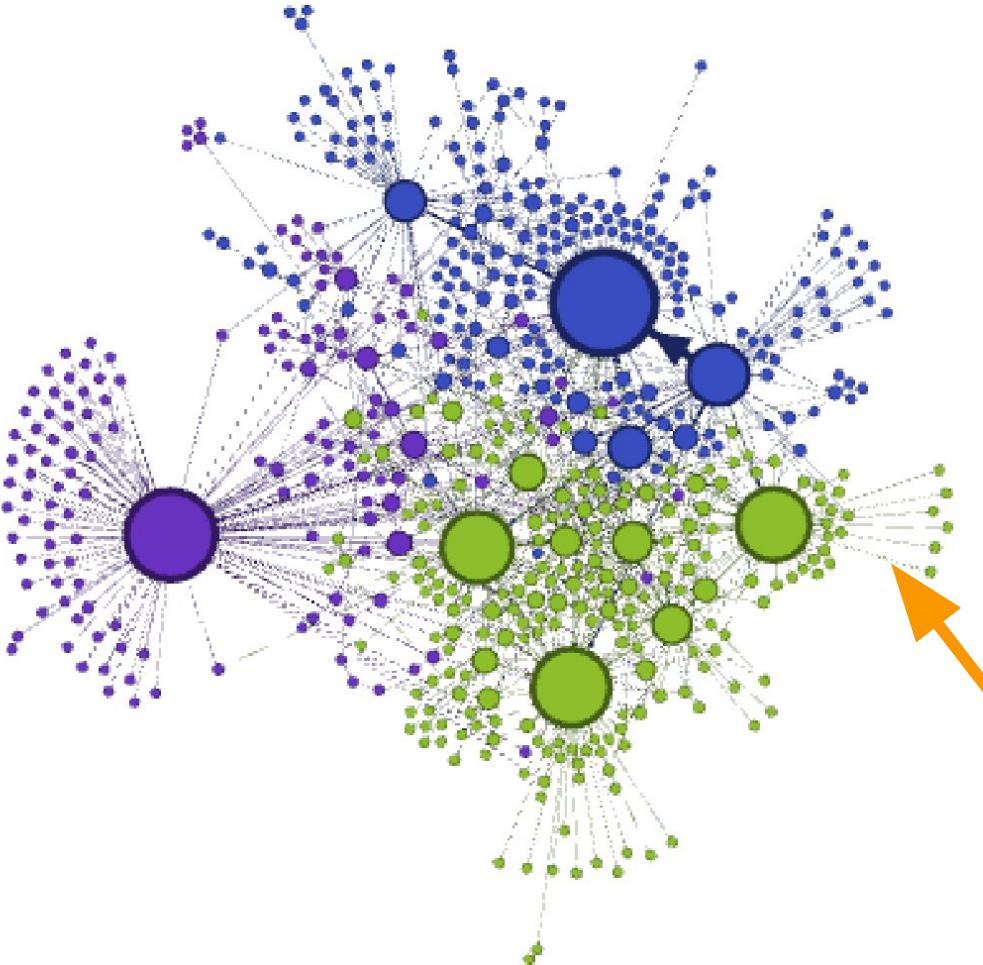
DÉTECTION COMMUNAU TAIRE DE LOUVAIN

Blondel, Vincent D., et al. "Déploiement rapide des communautés dans les grands réseaux." Journal de mécanique statistique : théorie et expérience 2008.10 (2008).









Best way to monetize an email list
Norman_drey · Yesterday at 4:02 PM · email list

Yesterday at 4:02 PM

INTRO
Regular Member

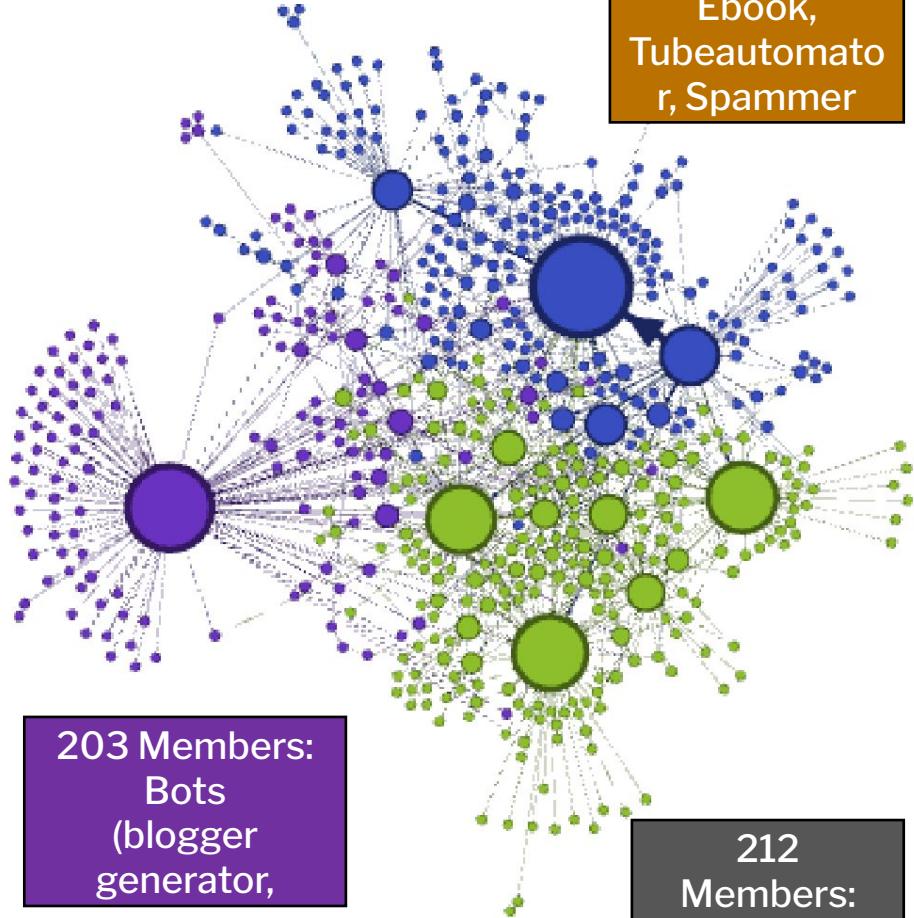
Joined: May 8, 2022
Messages: 336
Reaction score: 116

What's the best way to monetize an email list, I've been gathering emails from my entertainment blog, so far I've gotten over 100k list, I'm wondering if there could be an easier way to monetize it since I have no interest in paying for bulk email messaging service like mailchimp and others

Advertise on BHW



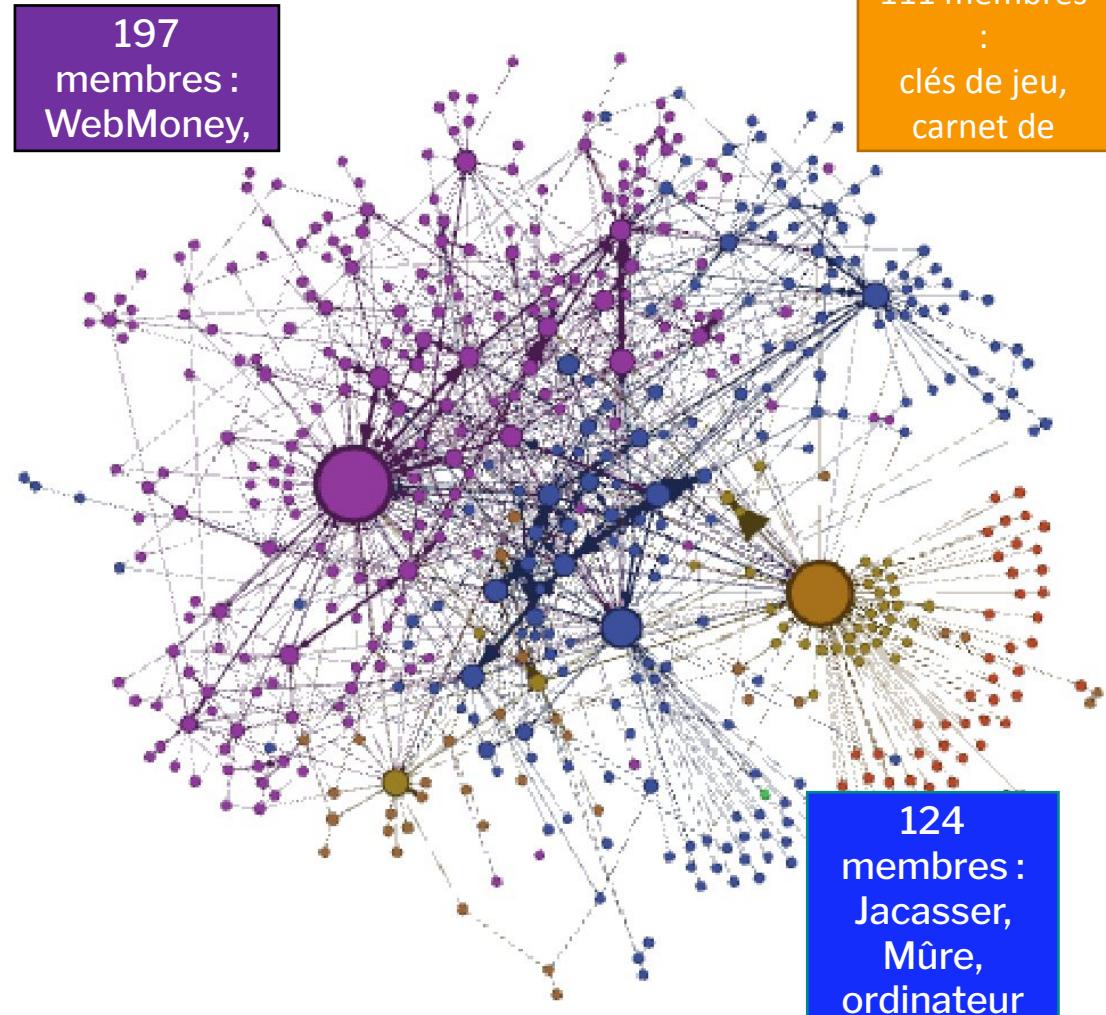
142 Members:
Ebook,
Tubeautomato
r, Spammer



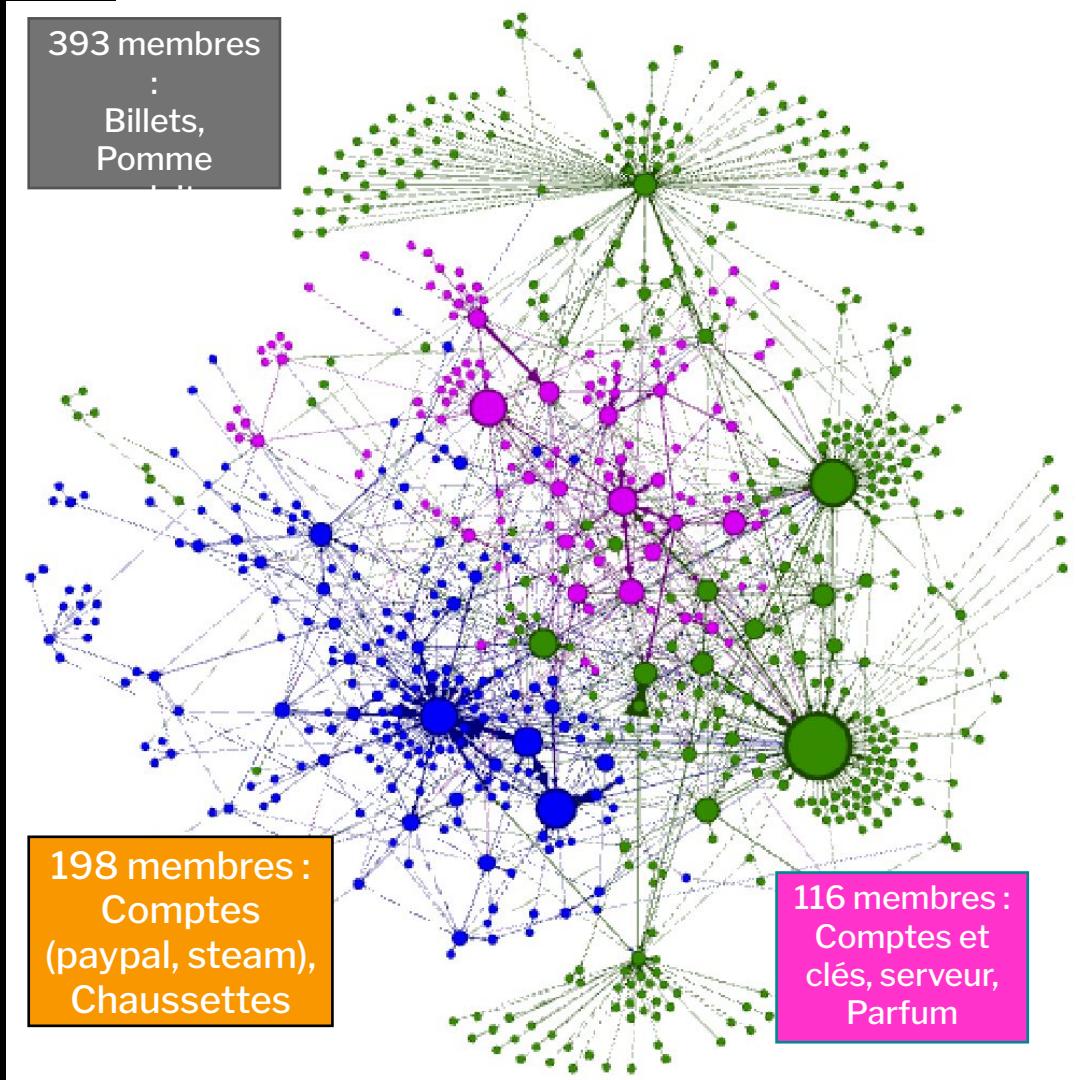
203 Members:
Bots
(blogger
generator,

212
Members:
Video upload,

Carders



L33tCrew



C #	BlackhatWorld		Carders		L33tCrew	
	Memb	Special topic	Memb	Special topic	Memb	Special topic
1	212	Video upload	800	Drugs	2348	Cardable shops
2	203	Blogger generator	527	Gametimecards	1696	Anonymity services
3	142	Ebook	375	WebMoney	1447	Apple devices
4	138	Account creators	352	Bots	1419	Crypter
5	104	Invites	311	Packstation	393	Tickets
6	99	Keyword stuffing	284	Fake packstation	198	Accounts
7	97	Xrumer	253	Video game	116	Perfume
8	93	Article generator	245	ATM skimmer	35	Trojans
9	90	Account creators	237	Cardable shops		
10	81	Torrents	231	VPN, WII		
11	79	Fantomaster	212	VPN		
12	77	Bulk email	197	Trojan		
13	60	Cloaking	111	Gamekeys		
14	59	Adsense	124	Jabber		
15	47	Cracked tools				
16	46	Stumblebot				
17	39	Tutorials				
18	16	Script				

A quoi ressemblent ces communautés ?

- Numéro Dunbar (150)
- Structure
- Mob-like vs Gang-like
- Les sujets
- Les sujets sont variés, pas uniformes, ce qui signifie que les communautés se spécialisent.

Motivation

- Comment la confiance évolue-t-elle ?
- Comment sont organisés les forums ?
- Qui est important ?
- Dirigeants, membres centraux
- Comment pouvons-nous les perturber ?
- Ou comment ne pouvons-nous pas?

Qui est important

Degree Centrality

- Raw number of connections
- Associated with higher trust

Betweenness Centrality

- Number of shortest paths that pass through the node
- More information

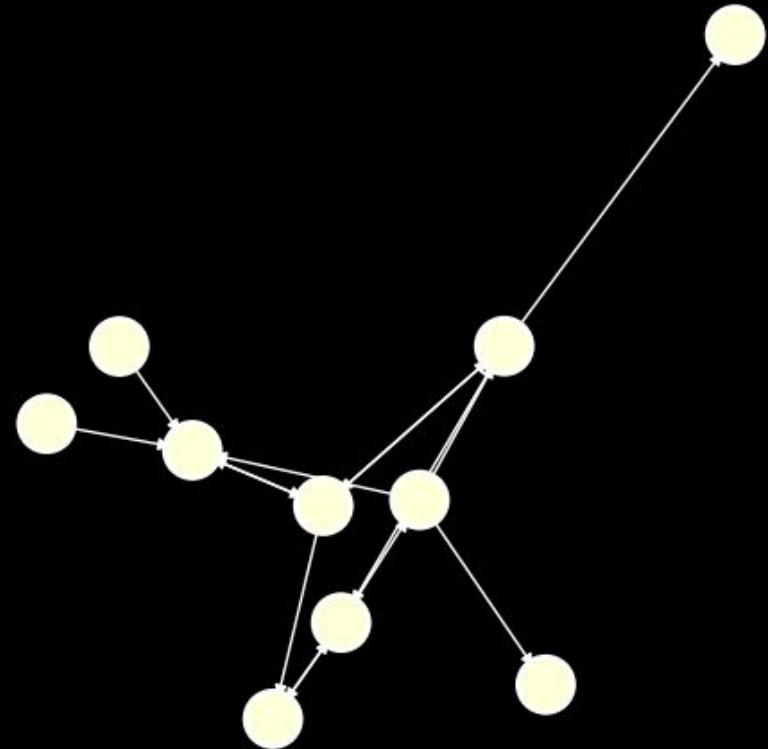
Closeness Centrality

- How far is this node from all other nodes
- Lowest transaction costs

Eigenvector Centrality

- How much influence does this node and its neighbors have

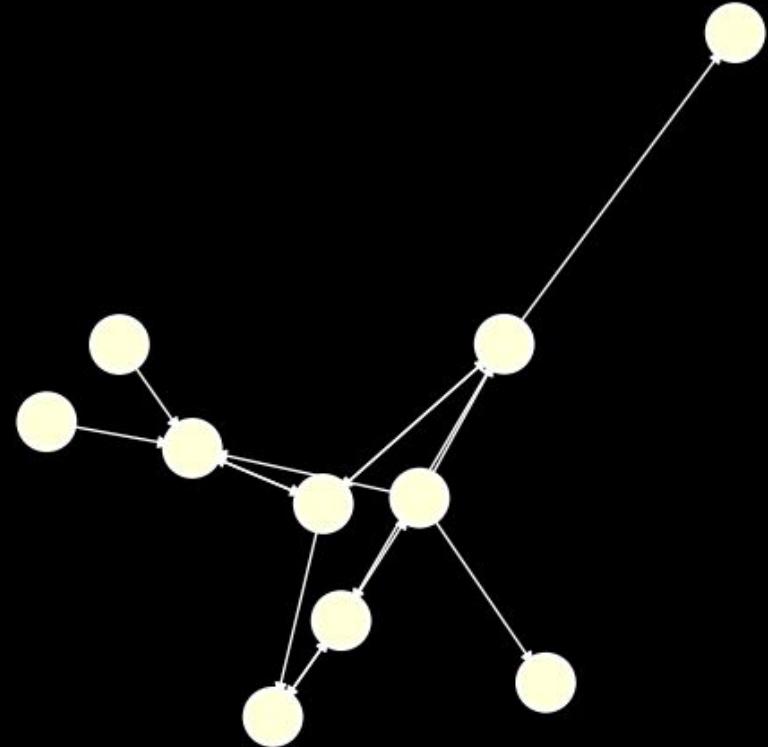
Construire un graphe social



Centralité des degrés

Nombre brut de connexions

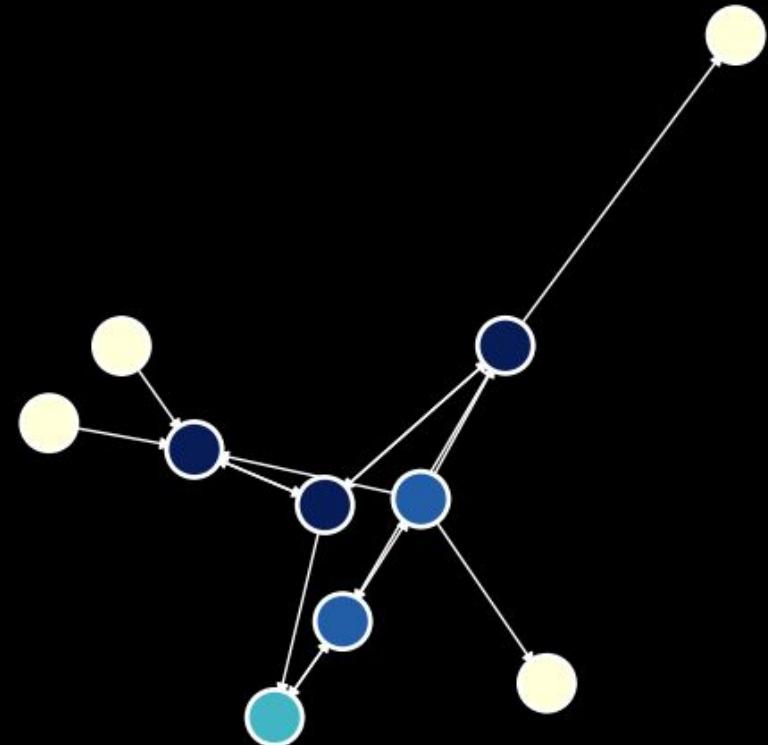
Associé à une plus grande
confiance



Centralité des degrés

Nombre brut de connexions

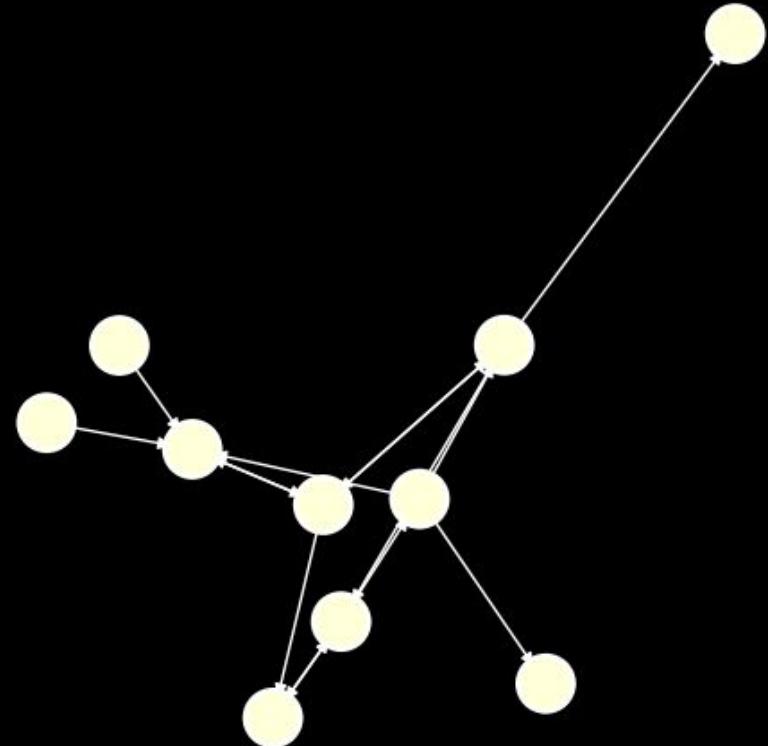
Associé à une plus grande
confiance



Centralité

Nombre de chemins les plus courts passant par le nœud

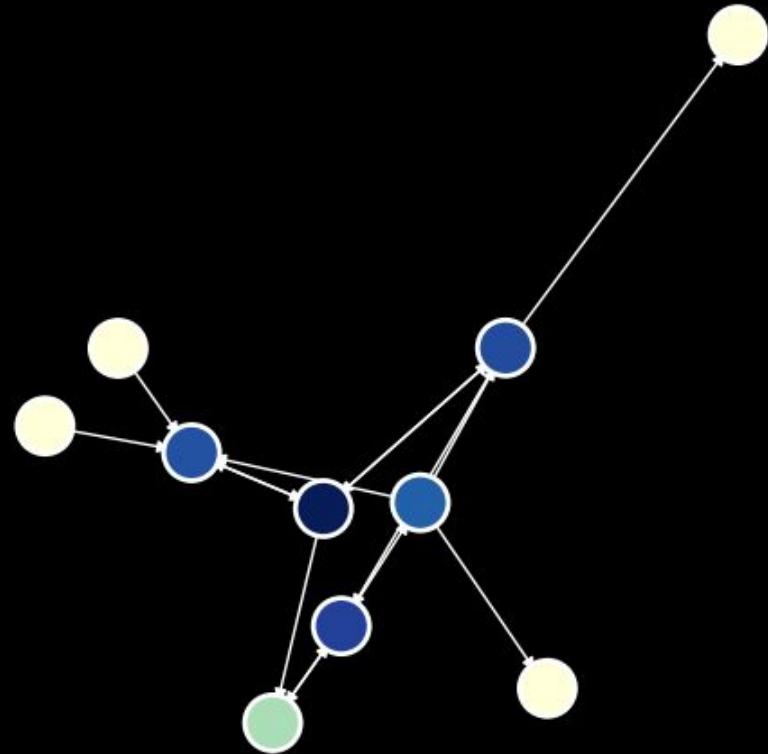
Plus d'information



Centralité

Nombre de chemins les plus courts passant par le nœud

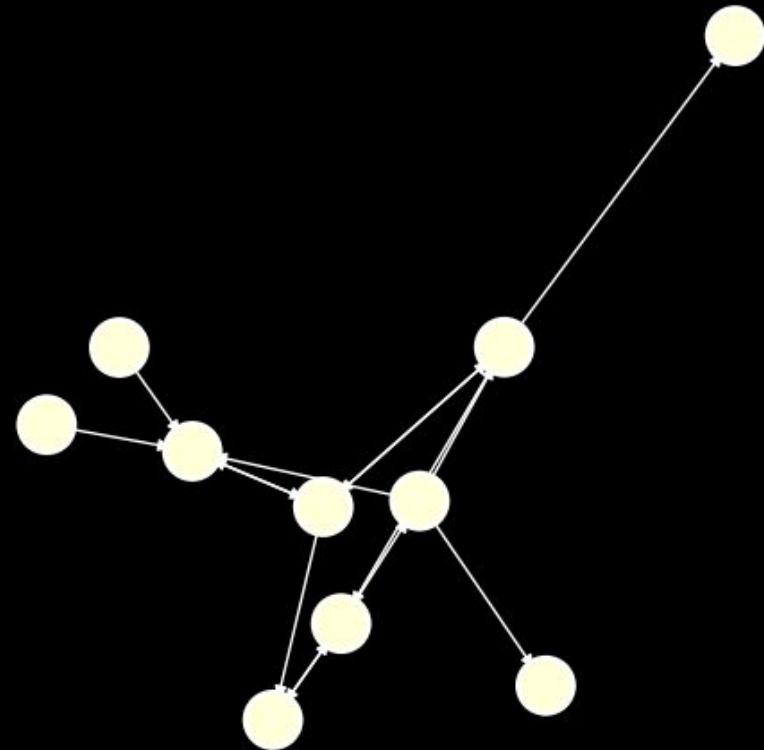
Plus d'information



Proximité Centralité

À quelle distance se trouve ce nœud de tous les autres nœuds

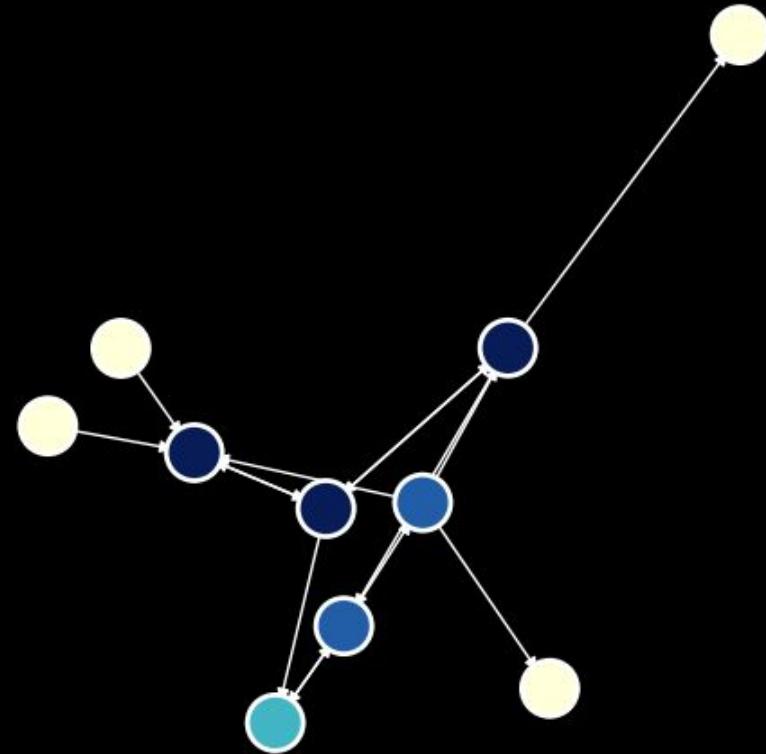
Coûts de transaction les plus bas



Proximité Centralité

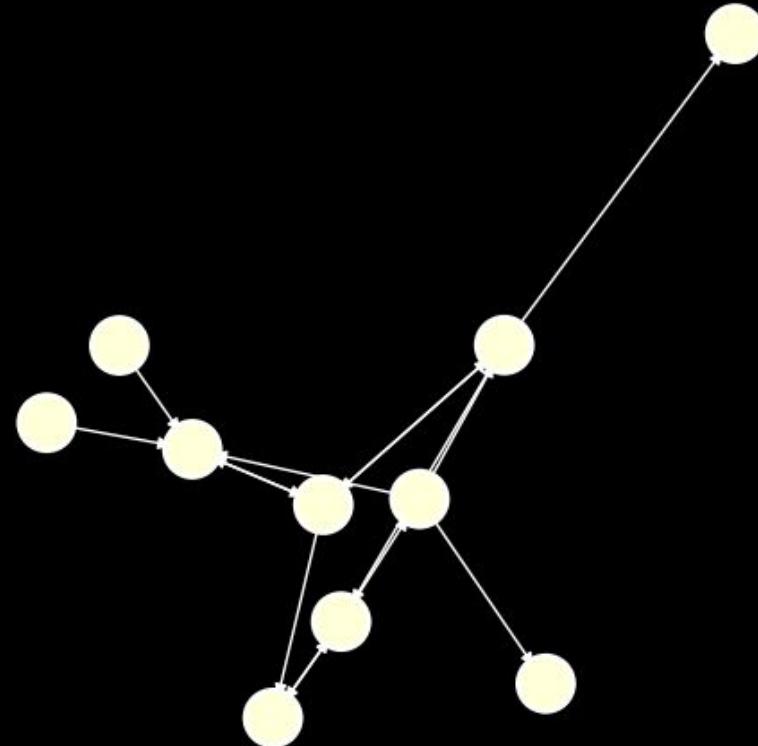
À quelle distance se trouve ce nœud de tous les autres nœuds

Coûts de transaction les plus bas



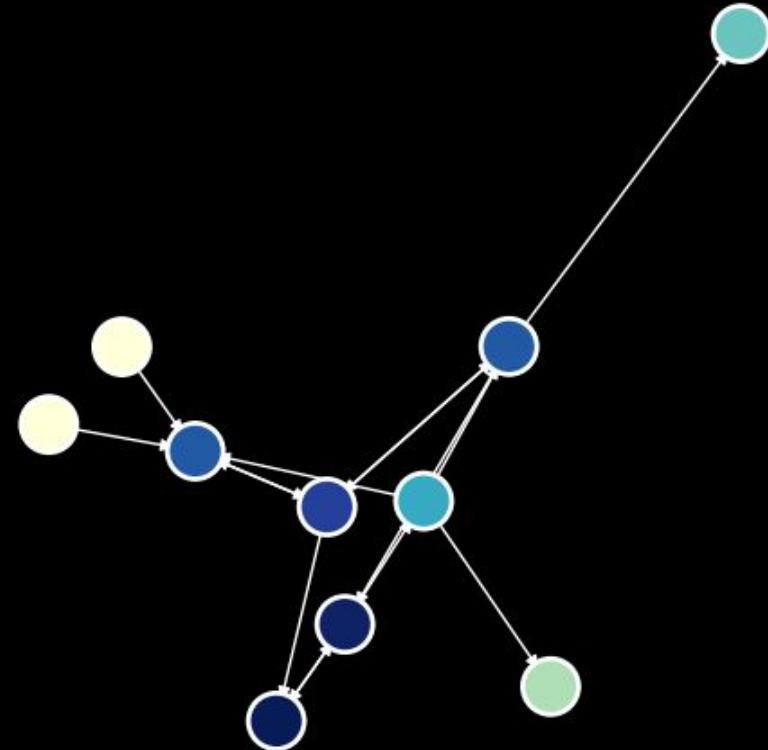
Centralité des vecteurs propres

Quelle est l'influence de ce
nœud et de ses voisins



Centralité des vecteurs propres

Quelle est l'influence de ce
nœud et de ses voisins



Résultats

- Sur BlackhatWorld – Tout est corrélé.
- Sur L33tCrew et Carders - Tout sauf la centralité de proximité est corrélé.
- Centralité de proximité - À quelle distance se trouve ce nœud de tous les autres nœuds

	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

Motivation

- Comment la confiance évolue-t-elle ?
- Comment sont organisés les forums ?
- Qui est important ?
- Dirigeants, membres centraux
- Comment pouvons-nous les perturber ?
- Ou comment ne pouvons-nous pas?

Interdiction

Comptes en double

Déchirure

Spammage

Que se passe-t-il lorsque des membres sont bannis ?

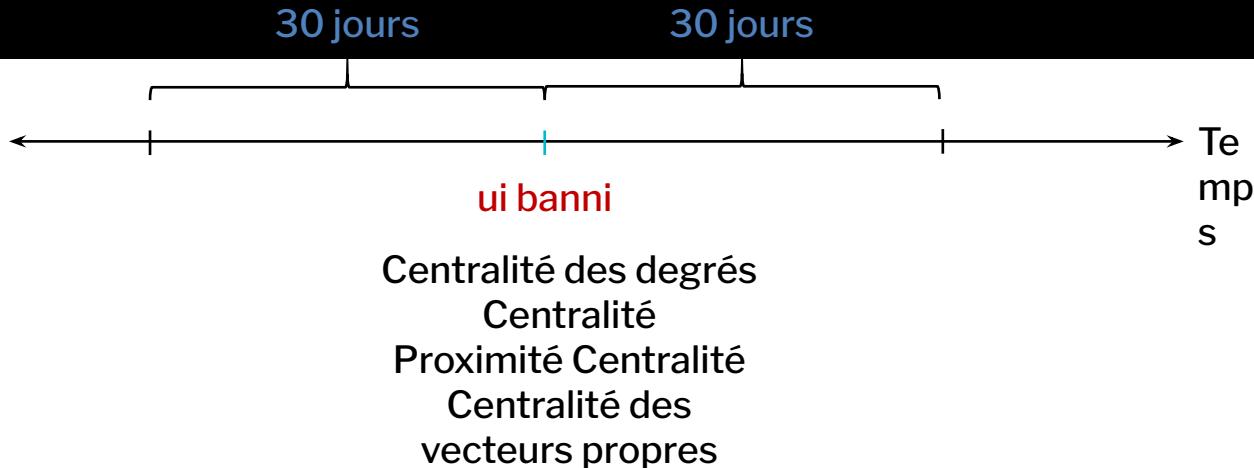


Que se passe-t-il lorsque des membres sont bannis ?



"Métriques du petit monde"
Coefficient de regroupement moyen
Longueur moyenne du chemin

Que se passe-t-il lorsque des membres sont bannis ?



Résultats

	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: $0.05 > * > 0.01 > ** > 0.001 > ***$

Résultats

	BlackhatWorld		Carders		L33tCrew	
CM	ΔACC	ΔAPL	ΔACC	ΔAPL	ΔACC	ΔAPL
Betweenness (B)	-0.39	0.32	-0.12***	-0.05*	La plupart des	0.01
Closeness (C)	0.07	0.12	-0.07**	-0.05*	-membres	0.11
Degree (D)	-0.18	0.22	-0.19***	-0.03	-bannis en	0.10
Eigenvector (E)	0.07	-0.12	-0.14***	-0.04	même temps	0.04

p-value: $0.05 > * > 0.01 > ** > 0.001 > ***$

Résultats

	BlackhatWorld		Carders		L33tCrew	
CM	ΔACC	ΔAPL	ΔACC	ΔAPL	ΔACC	ΔAPL
Betweenness (B)	-0.39	0.32	-0.12***	-0.05*	La plupart des	0.01
Closeness (C)	0.07	0.12	-0.07**	-0.05*	-membres	0.11
Degree (D)	-0.19***	0.22	-0.19***	-0.03	-bannis en	10
Eigenvector (E)	0.07	-0.12	-0.14***	-0.04	même temps	0.04

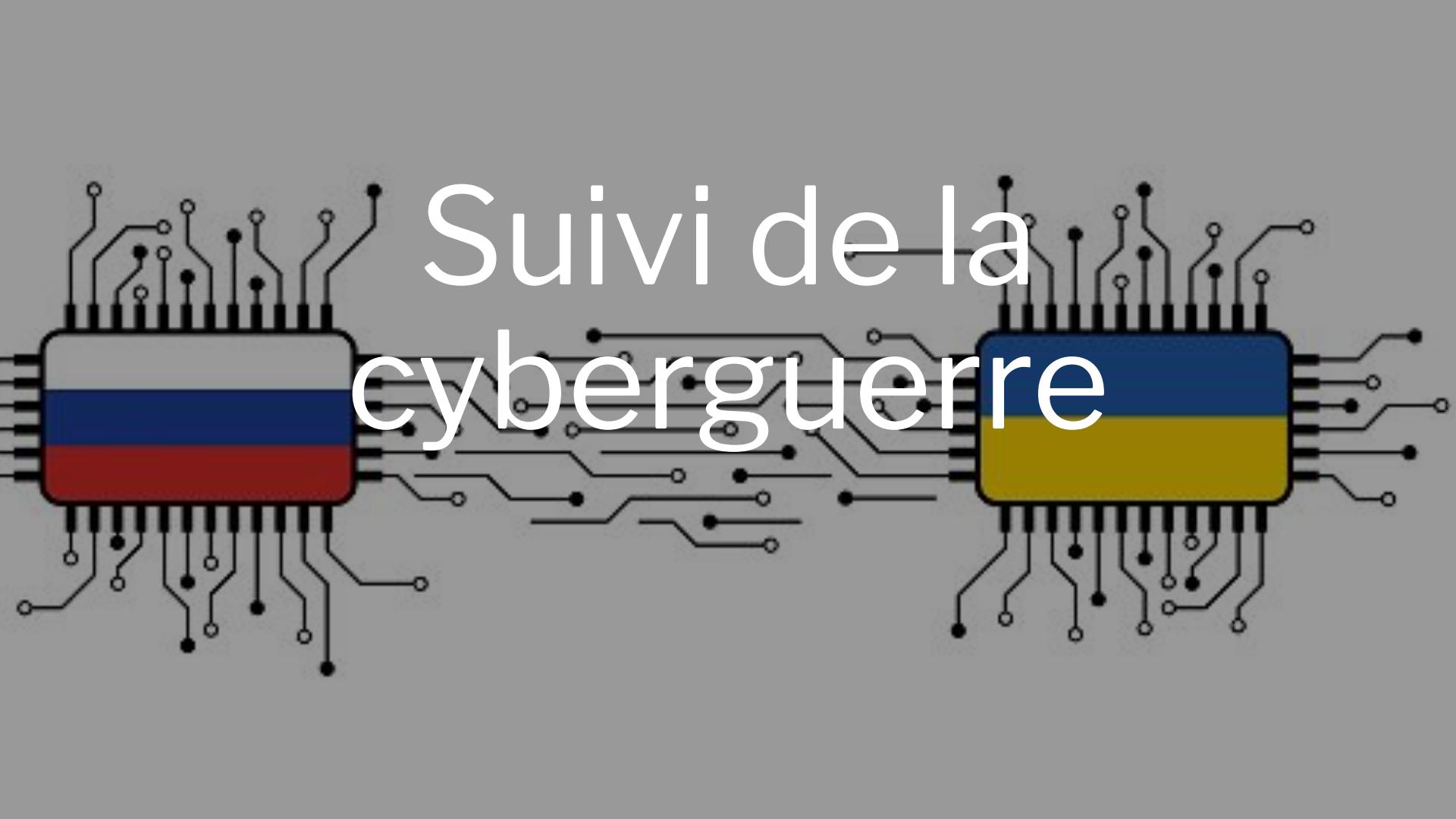
p-value: $0.05 > * > 0.01 > ** > 0.001 > ***$

Résultats

	BlackhatWorld		Carders		L33tCrew	
CM	ΔACC	ΔAPL	ΔACC	ΔAPL	ΔACC	ΔAPL
Betweenness (B)	-0.39	0.32	-0.12***	-0.05*	La plupart des	0.05***
Closeness (C)	0.07	0.12	-0.07**	-0.05*	-membres	-0.10***
Degree (D)	-0.15	0.22	-0.19***	-0.03	-bannis	-0.10***
Eigenvector (E)	0.07	-0.12	-0.14***	-0.04	en même temps	0.04***

p-value: $0.05 > * > 0.01 > ** > 0.001 > ***$

Les personnes bannies ne sont pas proches des autres nœuds.



Suivi de la
cyberguerre

ESET research

Follow ESETresearch

Feb 23 • 7 tweets • 2 min read

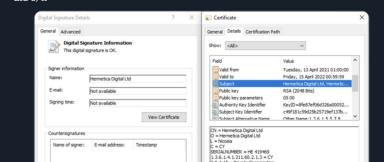
664 views

Bookmark Save as PDF + My Authors

Breaking. #ESETresearch discovered a new data wiper malware used in Ukraine today. ESET telemetry shows that it was installed on hundreds of machines in the country. This follows the DDoS attacks against several Ukrainian websites earlier today 1/n

We observed the first sample today around 14h52 UTC / 16h52 local time. The PE compilation timestamp of one of the sample is 2021-12-28, suggesting that the attack might have been in preparation for almost two months. 2/n

The Wiper binary is signed using a code signing certificate issued to Hermetica Digital Ltd 3/n



ESET Research

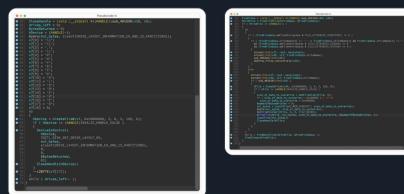
Follow ESETresearch

Mar 14, 2022 • 7 tweets • 4 min read

404 views

Bookmark Save as PDF + My Authors

#BREAKING #ESETresearch warns about the discovery of a 3rd destructive wiper deployed in Ukraine 🇺🇦. We first observed this new malware we call #CaddyWiper today around 9h38 UTC. 1/7



This new malware erases user data and partition information from attached drives. #ESET telemetry shows that it was seen on a few dozen systems in a limited number of organizations. 2/7

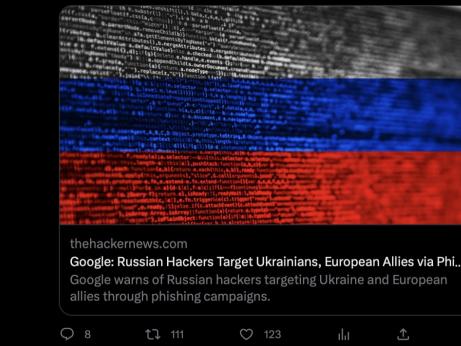
THN

The Hacker News @TheHackersNews • Mar 8, 2022

#Google warns that #Russian and Belarusian hackers are targeting Ukraine and European allies through phishing attacks.

Read details:thehackernews.com/2022/03/google...

#infosec #cybersecurity #hacking #malware



thehackernews.com
Google: Russian Hackers Target Ukrainians, European Allies via Phi...
Google warns of Russian hackers targeting Ukraine and European allies through phishing campaigns.

Attaques du début de la guerre depuis la Russie



Attaques du début de la guerre depuis l'Ukraine

Archives de la dégradation du Web

Scraped 4 archives de défiguration Web

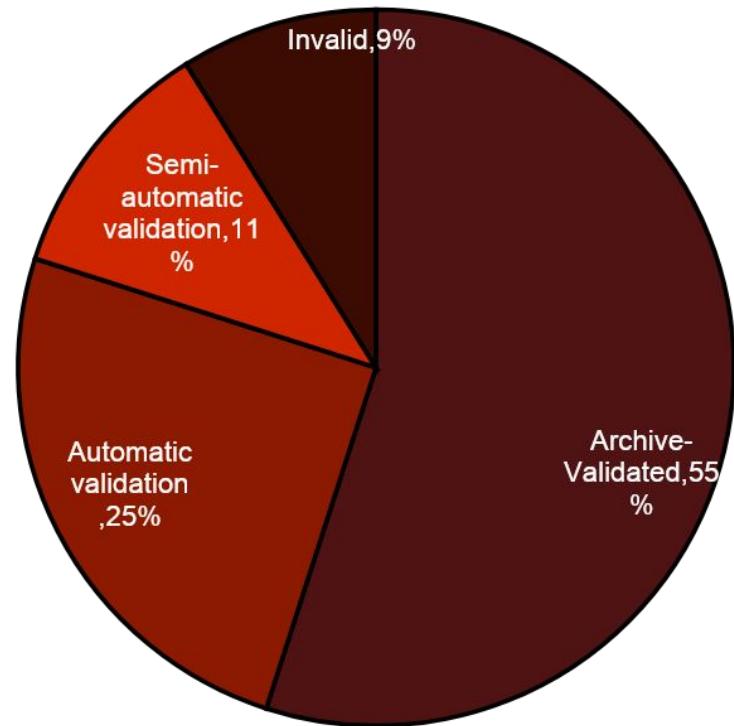
Filtré par pays par ccTLD (.ru/.ue) et IP

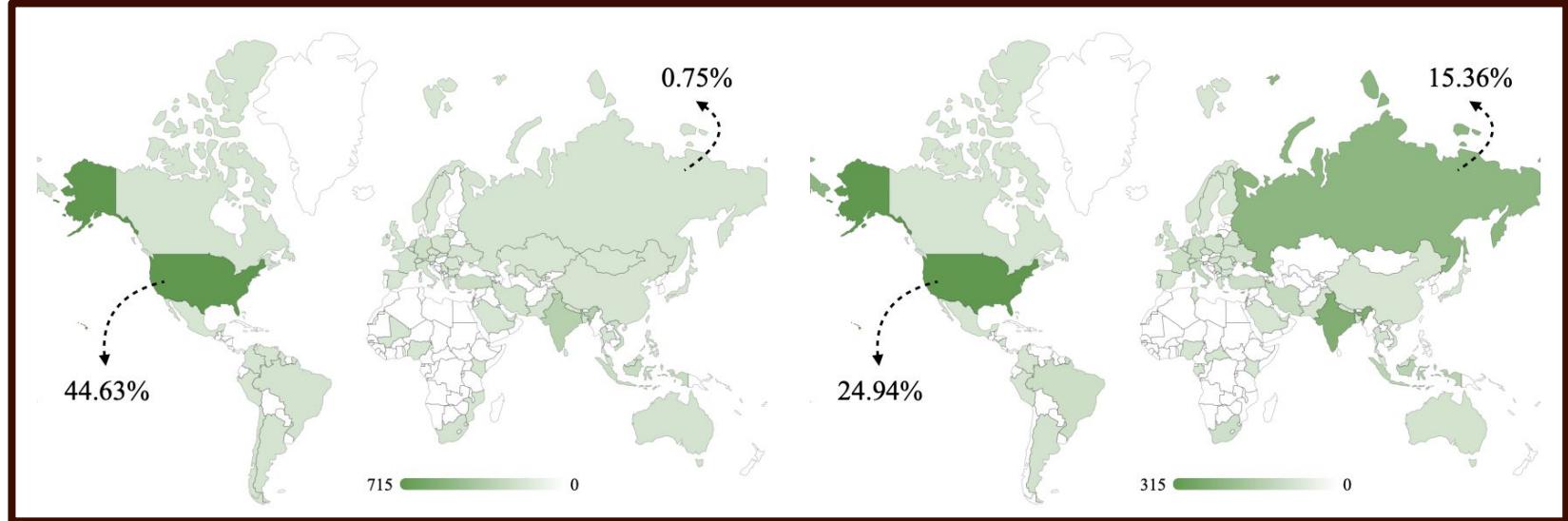
Doublons supprimés

Corriger les fautes de frappe dans le nom du "notificateur"

Validé les attaques

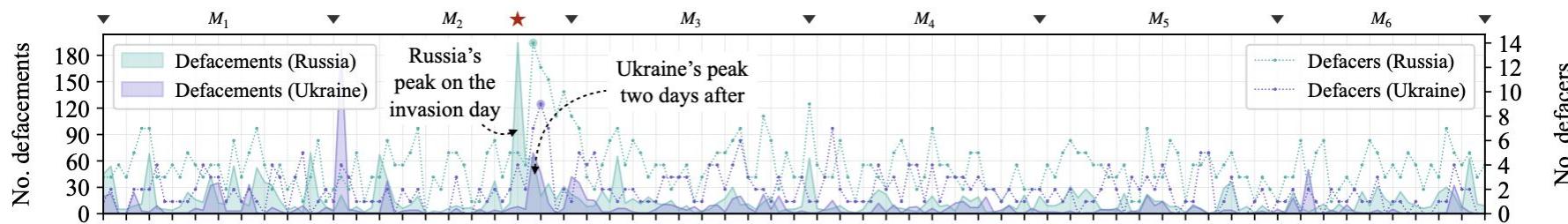
Validity of Submissions



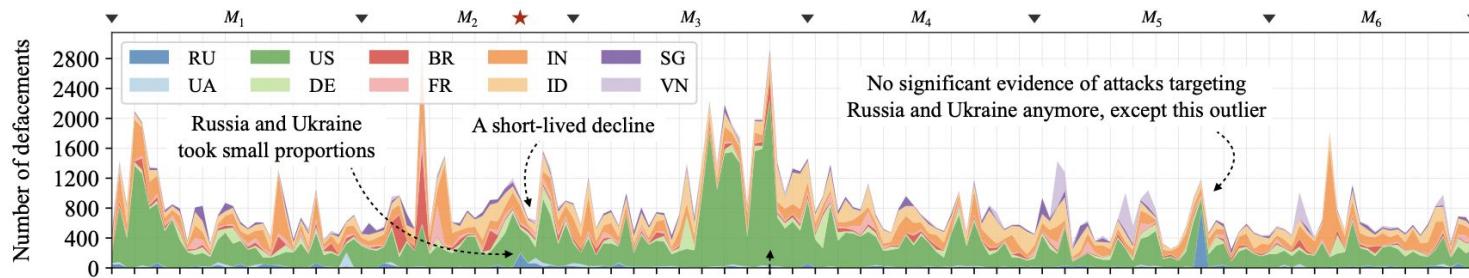


Jour avant vs jour de l'invasion

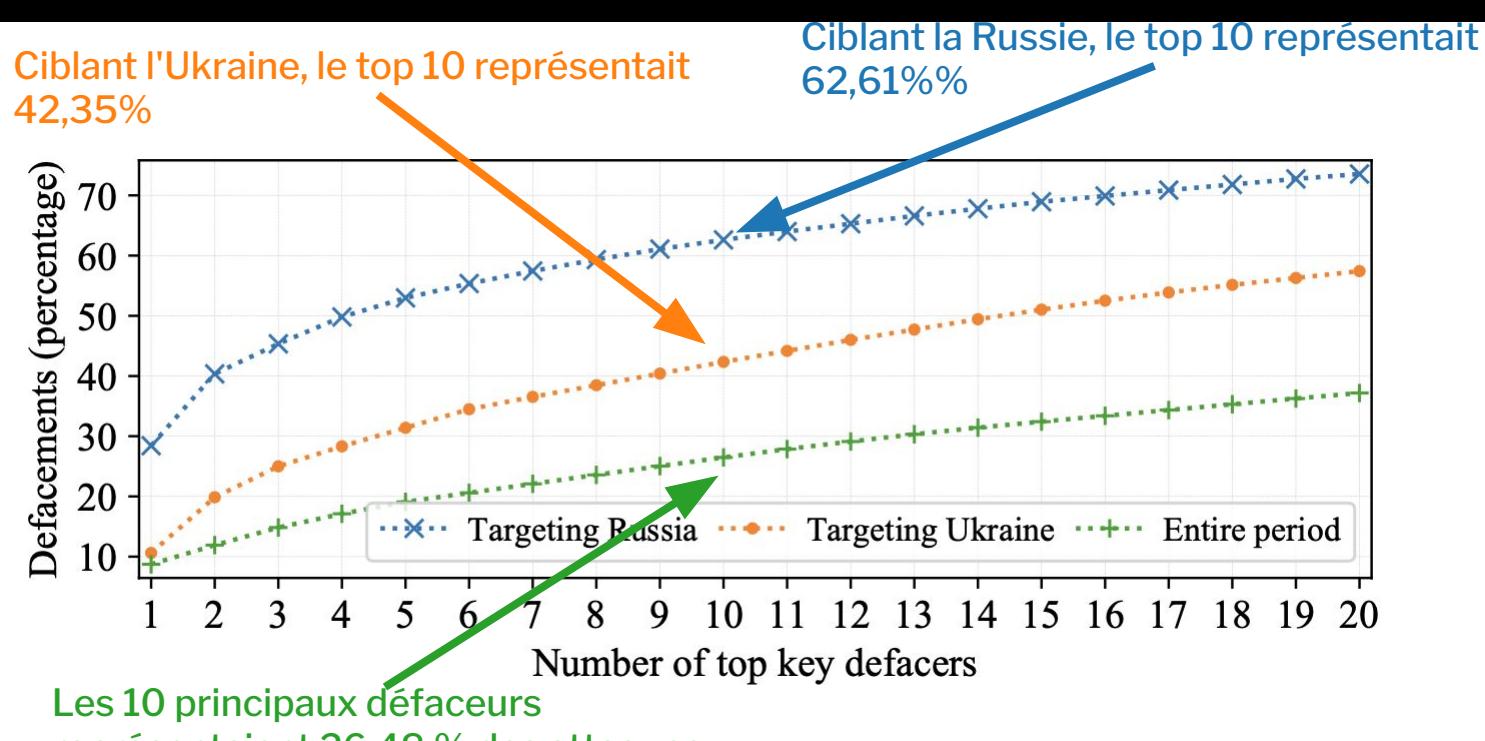
Attaques au fil du temps



Attaques dans le contexte mondial

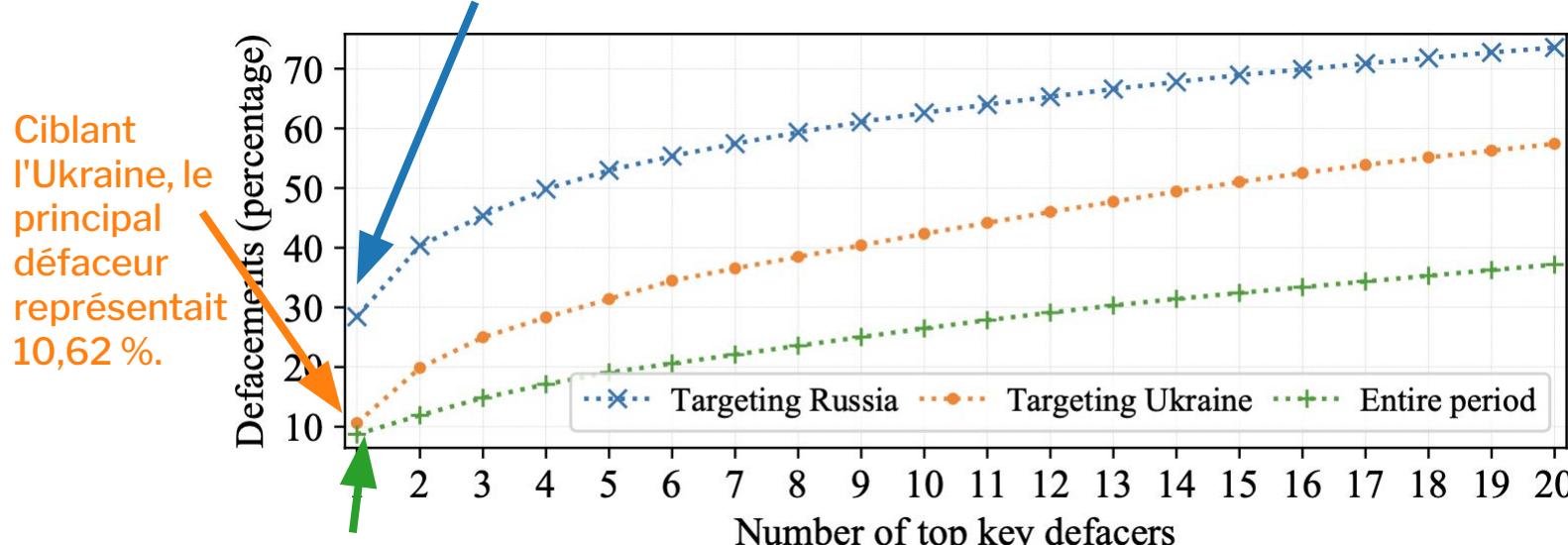


Concentration de défacants



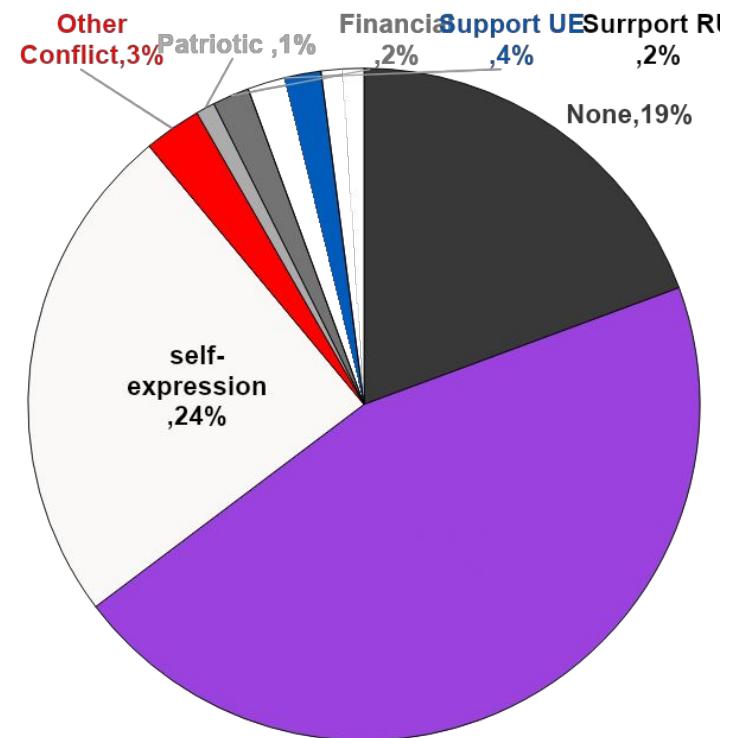
Concentration de défacants

Ciblant la Russie, le principal défaceur représentait 28,42 %.



Top defacer représentait 10 % des attaques

Motifs





Réseaux sociaux

- Des postes
- Les tendances
- Aime
- commentaires
- Votes
- Les sondages
- Reposts
- Popularité
- Suit
- Amis
- Vues
- Impressions

Comportement illicite sur les réseaux sociaux

Désinformation

Troll et faux comptes

Courrier indésirable

Piratage de compte

Traquer

Haine/harcèlement

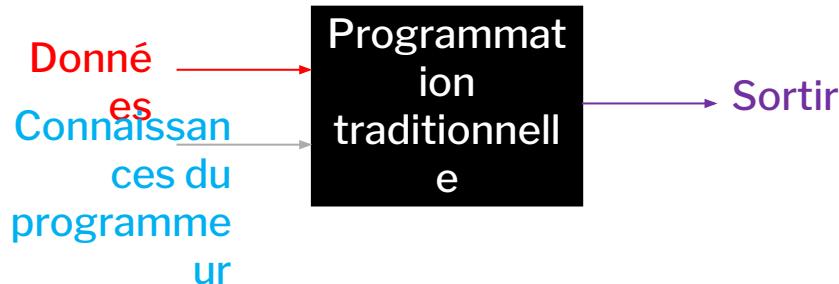
Inflation du contenu

...

Apprentissage automatique

Prédire un comportement illicite

Programmation traditionnelle



- Savoir quelque chose sur la tâche
- Indiquez explicitement à la machine comment le remplir
- Testez le modèle sur quelque chose que vous connaissez
- Déployez votre programme !



Exemple de programmation traditionnelle : détection de la parole moyenne

```
def detect_mean_speech(text):
    avec open('mean_words.txt', mode='r') comme f :
        mots_moyens = f.read().splitlines()

    pour mot dans text.split() :
        si mot dans mean_words :
            return True # Mot moyen trouvé :(
        return False # Aucun mot méchant n'a été trouvé :)
```

Tu es un visage stupide et stupide !

Tu ressembles à une grenouille !



Exemple de programmation traditionnelle : détection de la parole moyenne

```
def detect_mean_speech(text):
    avec open('mean_words.txt', mode='r') comme f :
mots_moyens = f.read().splitlines()

pour mot dans text.split() :
si mot dans mean_words :
return True # Mot moyen trouvé :(
return False # Aucun mot méchant n'a été trouvé :)
```

moyen
ne
stupid
e
bête
bizarre
abrut
idiot

mots_moyens.txt

Tu es un visage stupide et stupide !

Tu ressembles à une grenouille !



Exemple de programmation traditionnelle : détection de la parole moyenne

```
def detect_mean_speech(text):
    avec open('mean_words.txt', mode='r') comme f :
        mots_moyens = f.read().splitlines()

    pour mot dans text.split() :
        si mot dans mean_words :
            return True # Mot moyen trouvé :(
    return False # Aucun mot méchant n'a été trouvé :)
```

moyen
ne
stupid
e
bête
bizarre
abrut
idiot

mot_moyens.txt

Tu es un visage stupide et stupide !

Tu ressembles à une grenouille !



Exemple de programmation traditionnelle : détection de la parole moyenne

```
def detect_mean_speech(text):
    avec open('mean_words.txt', mode='r') comme f :
        mots_moyens = f.read().splitlines()

    pour mot dans text.split() :
        si mot dans mean_words :
            return True # Mot moyen trouvé :(
        return False # Aucun mot méchant n'a été trouvé :)
```

moyen
ne
stupid
e
bête
bizarre
abrutil
mots_moyens.txt
idiot

Tu es un visage stupide et stupide !

Tu ressembles à une grenouille !



Exemple de programmation traditionnelle pour détection de parole méchante



Twitter Safety ✨ @TwitterSafety · 17h

To quantify hate speech, Twitter & Sprinklr start with 300 of the most common English-language slurs. We count not only how often they're tweeted but how often they're seen (impressions).

Our models score slur Tweets on “toxicity,” the likelihood that they constitute hate speech



42



117



786



89.7K



```
def detect_mean_speech(text):
    avec open('mean_words.txt') comme f:
        mots_moyens = f.read().splitlines()
```

```
pour mot dans text.split():
    si mot dans mean_words :
        return True # Mot moyen trouvé :(
    return False # Aucun mot méchant n'a été trouvé :)
```

beta
bizarre
abrut
idiot
mots_moyens.txt

Tu es un visage stupide et stupide !

Tu ressembles à une grenouille !

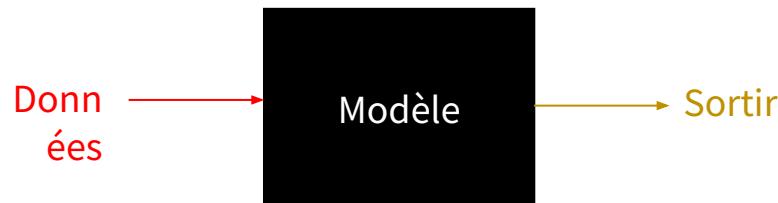
Apprentissage automatique (ML)

La capacité d'apprendre sans être explicitement programmé



Exploration de données : obtenez des données dont vous savez quelque chose

Entraîner le modèle : « enseignez » à la machine ces données

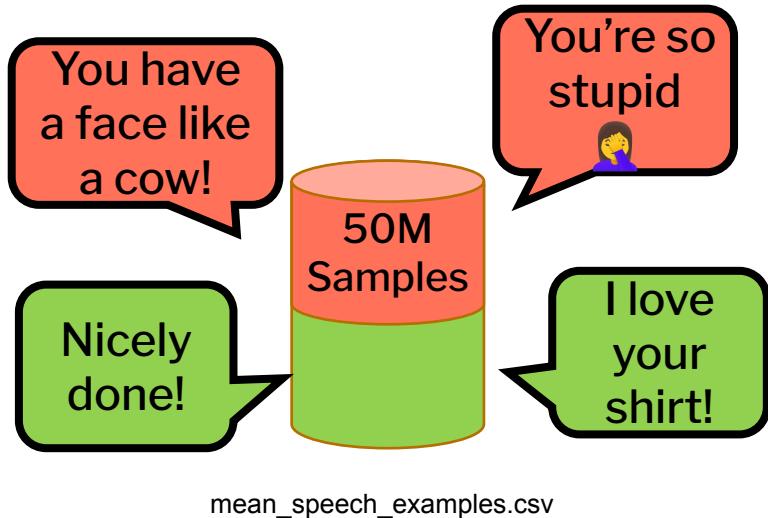


Testez le modèle : "Vérifiez" que la machine sur quelque chose que vous connaissez

Déployez le programme ! : utilisez les apprentissages pour classer/prédire sur de nouvelles données

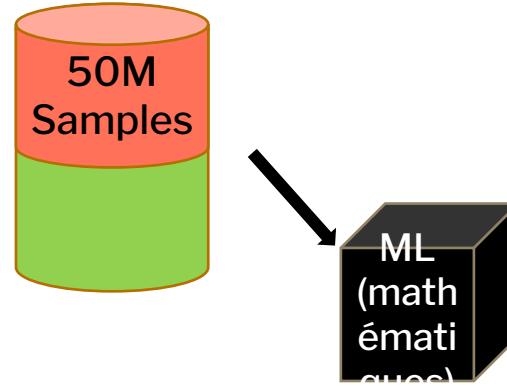
Exemple d'apprentissage automatique: détection de la parole moyenne

```
def detect_hate_speech_with_ML(text):
    df = pd.read_csv('hate_speech_examples.csv')
    # transformer le texte en vecteur
    vectoriseur = TfidfVectorizer()
    data = vectorizer.fit_transform(df['text'])
    labels = dataset['classe'].astype(int)
    # apprentissage automatique
    modèle = RandomForest().fit(données,étiquettes)
    si modèle.predict(text):
        return True # mot moyen trouvé ! :(
    autre:
        return False # aucun mot moyen n'a été trouvé :)
```



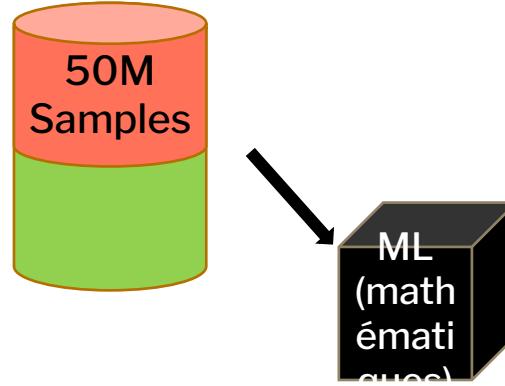
Exemple d'apprentissage automatique : détection de la parole moyenne

```
def detect_hate_speech_with_ML(text):
    df = pd.read_csv('hate_speech_examples.csv')
    # transformer le texte en vecteur
    vectoriseur = TfidfVectorizer()
    data = vectorizer.fit_transform(df['text'])
    labels = dataset['classe'].astype(int)
    # apprentissage automatique
    modèle = RandomForest().fit(données, étiquettes)
    si modèle.predict(text):
        return True # mot moyen trouvé ! :(
    autre:
        return False # aucun mot moyen n'a été trouvé :)
```



Exemple d'apprentissage automatique : détection de la parole moyenne

```
def detect_hate_speech_with_ML(text):
    df = pd.read_csv('hate_speech_examples.csv')
    # transformer le texte en vecteur
    vectoriseur = TfidfVectorizer()
    data = vectorizer.fit_transform(df['text'])
    labels = dataset['classe'].astype(int)
    # apprentissage automatique
    modèle = RandomForest().fit(données, étiquettes)
    si modèle.predict(text):
        return True # mot moyen trouvé ! :(
    autre:
        return False # aucun mot moyen n'a été trouvé :)
```

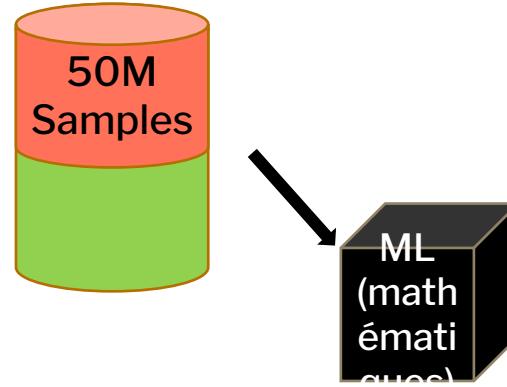


Tu es un
visage stupide
et stupide !

Tu ressembles
à une
grenouille !

Exemple d'apprentissage automatique : détection de la parole moyenne

```
def detect_hate_speech_with_ML(text):
    df = pd.read_csv('hate_speech_examples.csv')
    # transformer le texte en vecteur
    vectoriseur = TfidfVectorizer()
    data = vectorizer.fit_transform(df['text'])
    labels = dataset['classe'].astype(int)
    # apprentissage automatique
    modèle = RandomForest().fit(données, étiquettes)
    si model.predict(text):
        return True # mot moyen trouvé ! :(
    autre:
        return False # aucun mot moyen n'a été trouvé :)
```



Tu es un
visage stupide
et stupide !

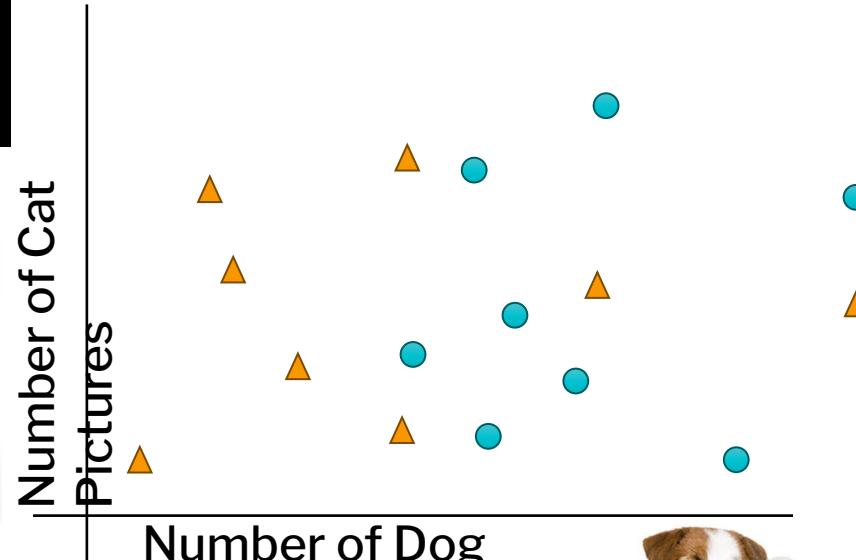
Tu ressembles
à une
grenouille !

Un classificateur simple : kNN

Question:
Un utilisateur de Facebook a-t-il un chat ?



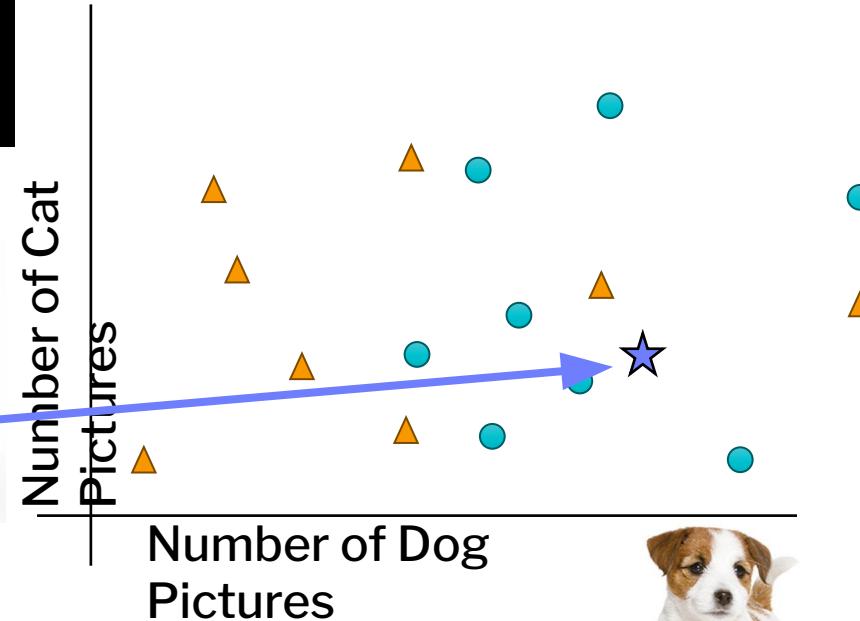
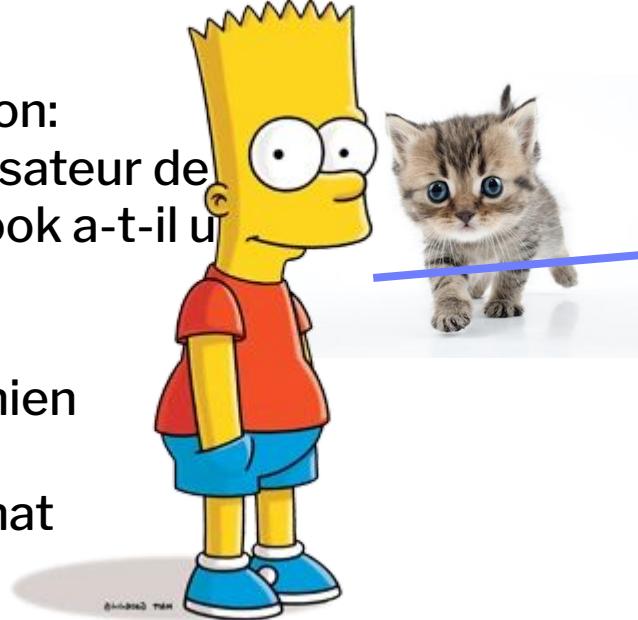
- A un chien
- ▲ A un chat



Un classificateur simple : kNN

Question:
Un utilisateur de Facebook a-t-il un chat ?

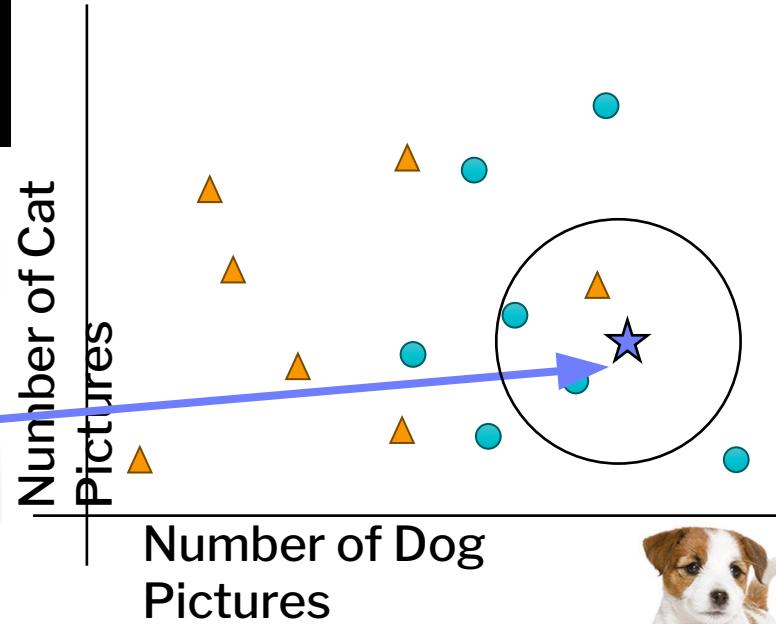
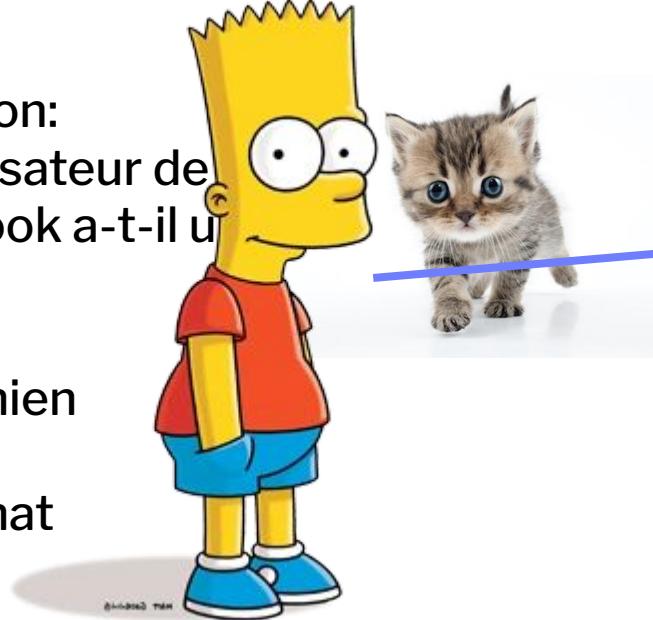
- A un chien
- ▲ A un chat



Un classificateur simple : kNN

Question:
Un utilisateur de Facebook a-t-il un chat ?

- A un chien
- ▲ A un chat

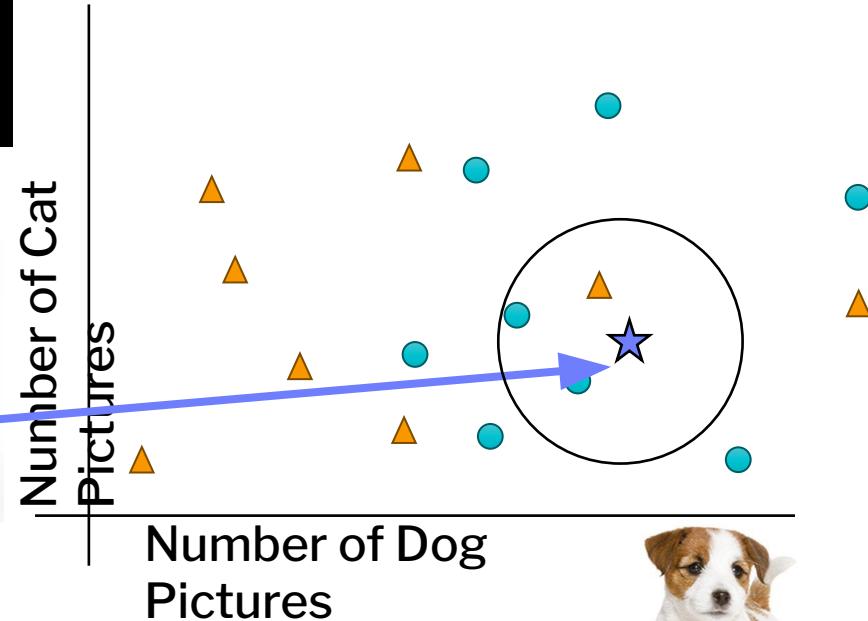


Un classificateur simple : kNN

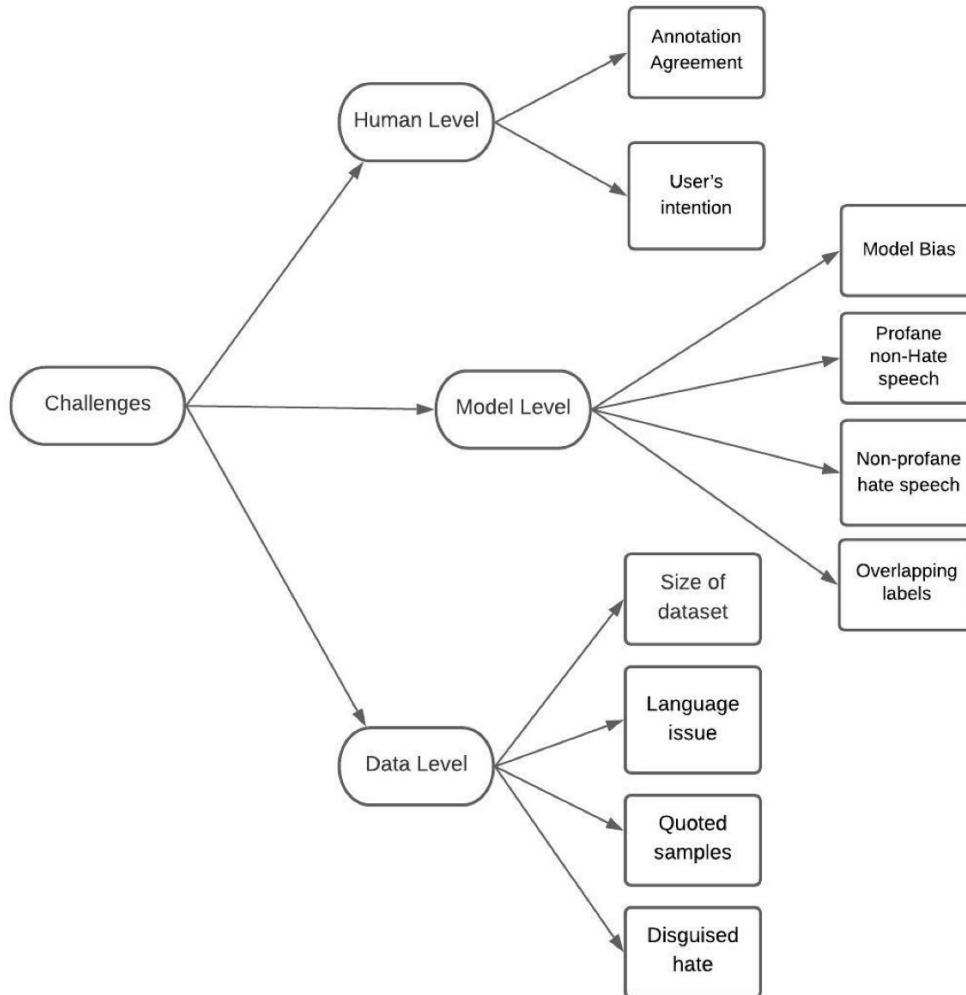
Question:
Un utilisateur de Facebook a-t-il un chat ?



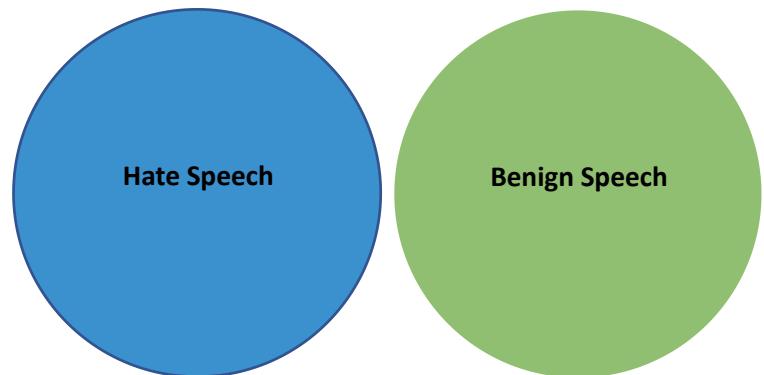
- A un chien
- ▲ A un chat



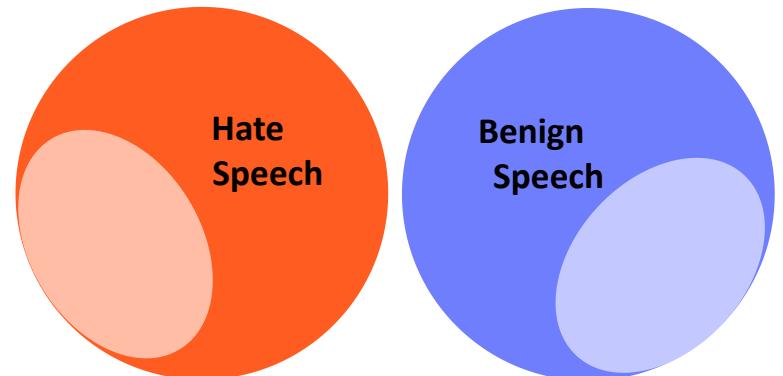
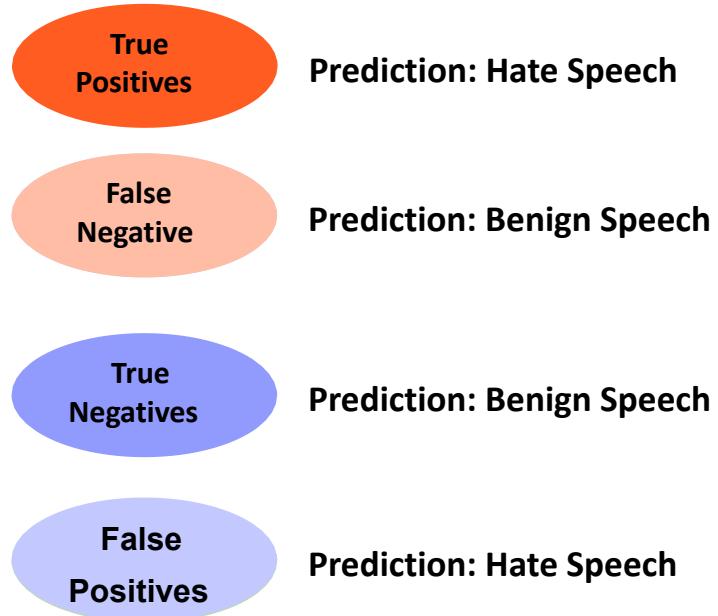
Défis de détection des discours haineux



Erreur du taux de base / erreur du procureur



Erreur du taux de base / erreur du procureur



Erreur du taux de base / erreur du procureur

True
Positives

Prediction: Hate Speech

False
Negative

Prediction: Benign Speech

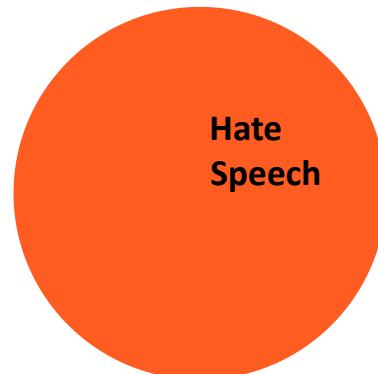
True
Negatives

Prediction: Benign Speech

False
Positives

Prediction: Hate Speech

- Notre classificateur :
- Taux de vrais positifs de 100 %
- Pas de faux négatifs, ne rate jamais le discours de haine
- Si le message contient un discours de haine, le classificateur dira OUI
- Taux de faux positifs de 5 %
- 5 fois sur 100 un message bénin est marqué comme discours de haine



Erreur du taux de base / erreur du procureur

True
Positives

Prediction: Hate Spe

False
Negative

Prediction: Benign Speech

True
Negatives

Prediction: Benign Speech

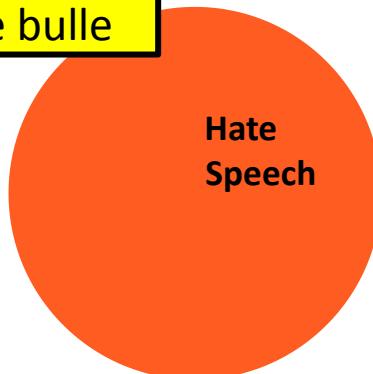
False
Positives

Prediction: Hate Speech

- Notre classificateur :
- Taux de vrais positifs de 100 %
- Pas de faux négatifs, ne rate jamais le discours de haine
- Si le message contient un discours de haine, le classificateur dira OUI

A quoi ça ressemble:
Super! La bulle FP est
vraiment petite ! Et il n'y
a pas du tout de bulle

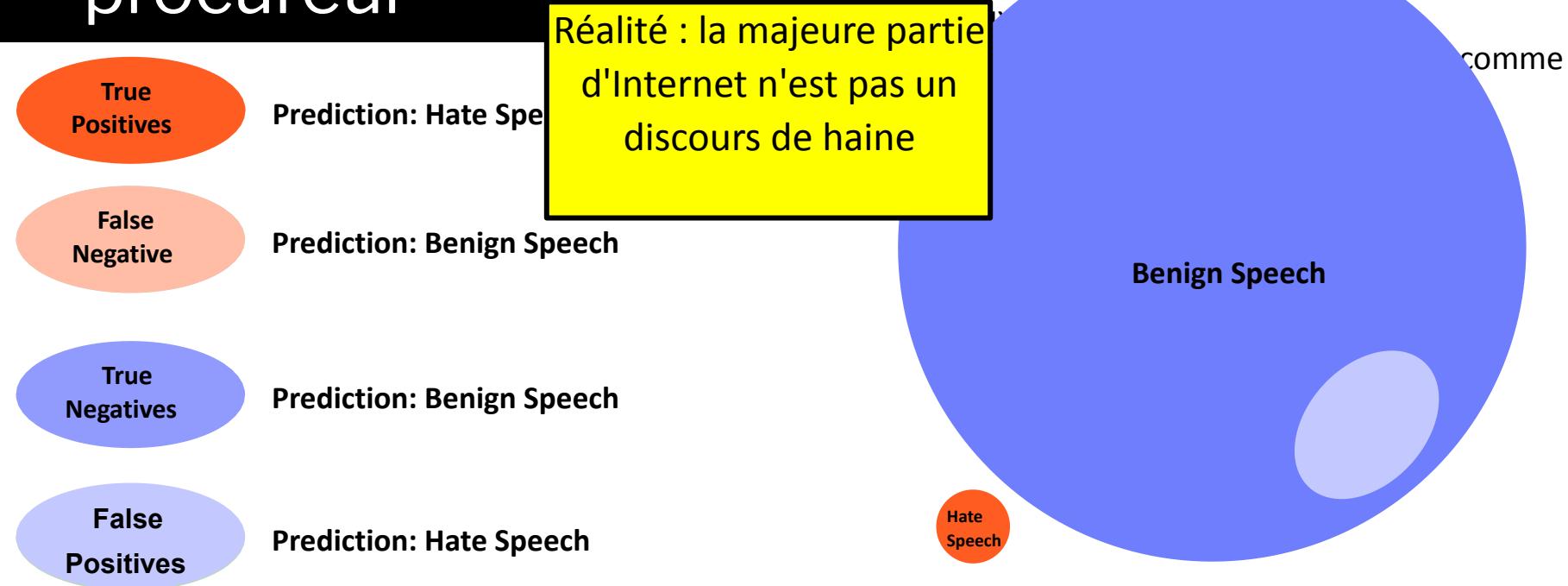
FN !



vraux positifs de 5 %
00 un message bénin est marqué comme
haine

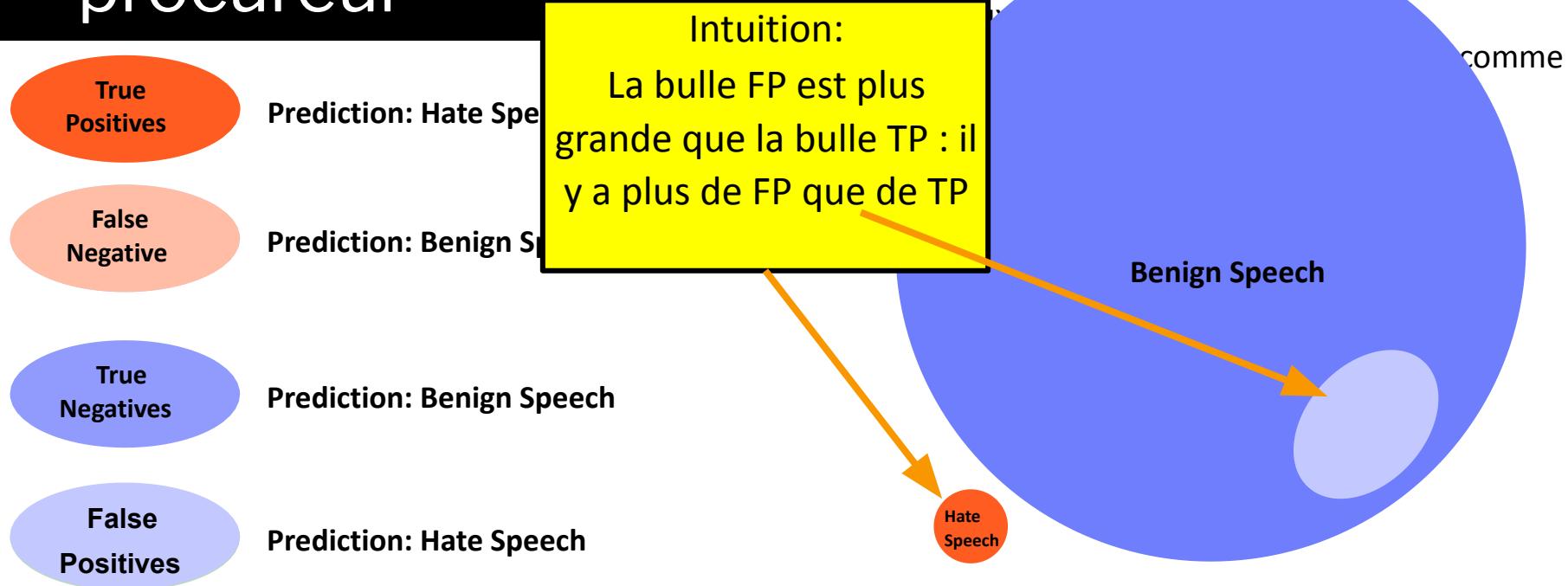
Erreur du taux de base / erreur du procureur

- Notre classificateur :
- Taux de vrais positifs de 100 %
- Pas de faux négatifs, ne rate jamais le discours de haine
- Si le message contient un discours de haine, le classificateur dit : « Oui »



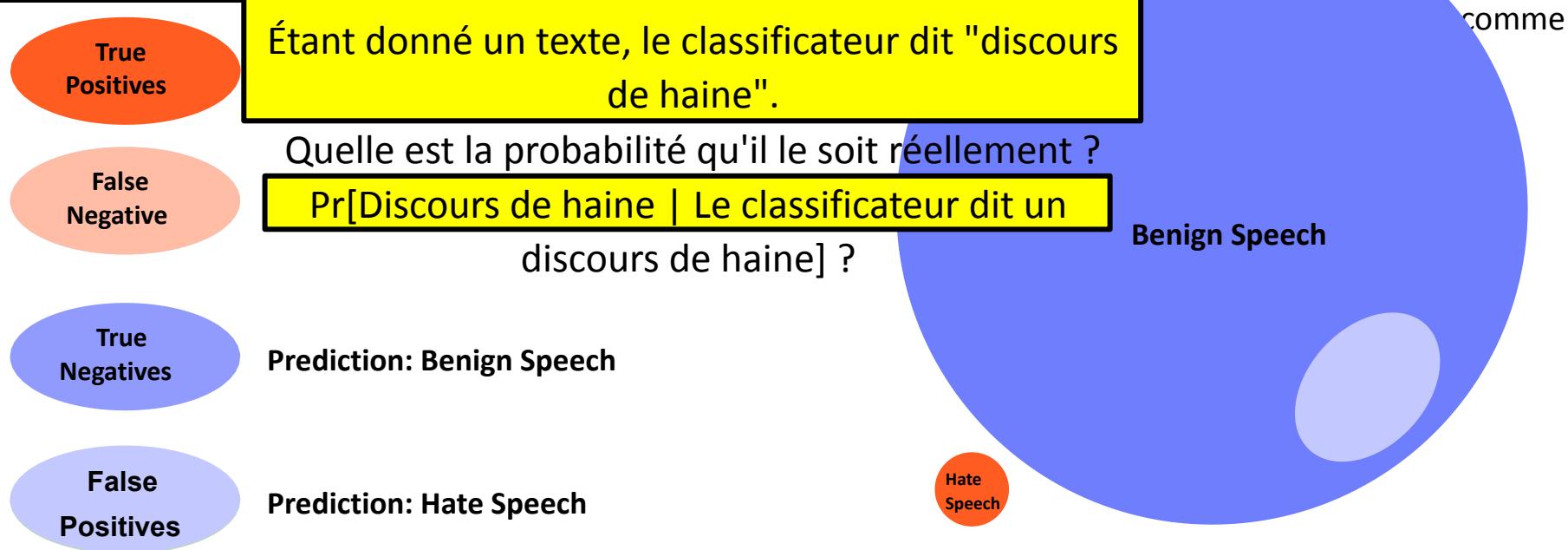
Erreur du taux de base / erreur du procureur

- Notre classificateur :
- Taux de vrais positifs de 100 %
- Pas de faux négatifs, ne rate jamais le discours de haine
- Si le message contient un discours de haine, le classificateur dit toujours que c'est du haine



Erreur du taux de base / erreur du procureur

- Notre classificateur :
- Taux de vrais positifs de 100 %
- Pas de faux négatifs, ne rate jamais le discours de haine
- Si le message contient un discours de haine, le classificateur dit



Erreur du taux de base / erreur du procureur

Notre classificateur :

Taux de vrais positifs de 100 %

Zéro FN

Taux de faux positifs de 5 %

5 fois sur 100 un message bénin est marqué comme discours de haine

"Réalité": 1 message sur 1000 est un discours de haine

Les maths:

Étant donné un texte, le classificateur dit "discours de haine".

Quelle est la probabilité qu'il le soit réellement ?

$\Pr[\text{Discours de haine} \mid \text{Le classificateur dit un discours de haine}]$?

Qui connaît assez le théorème de Bayes pour comprendre cela ?

Erreur du taux de base / erreur du procureur

Hints:

Bayes Theorem:

$$p[A|B] = \frac{p[B|A] p[A]}{p[B]}$$

Law of total probability:

$$P[A] = \sum_n p[A|B_n]p(B_n)$$

$$P[A] = p[A|B_1]p(B_1) + p(A|B_2)p(B_2)$$

$$\begin{aligned} B_1 &= \text{Hate Speech} \\ B_2 &= \text{Benign} \end{aligned}$$

Notre classificateur :

Taux de vrais positifs de 100 %

Zéro FN

Taux de faux positifs de 5 %

5 fois sur 100 un message bénin est marqué comme discours de haine

"Réalité": 1 message sur 1000 est un discours de haine

Les maths:

Étant donné un texte, le classificateur dit "discours de haine".

Quelle est la probabilité qu'il le soit réellement ?

Pr[Discours de haine | Le classificateur dit un
discours de haine] ?

Qui connaît assez le théorème de Bayes pour comprendre cela ?

Erreur du taux de base / erreur du procureur

Notre classificateur :

Taux de vrais positifs de 100 %

Zéro FN

Taux de faux positifs de 5 %

5 fois sur 100 un message bénin est marqué comme discours de haine

"Réalité": 1 message sur 1000 est un discours de haine

Pr[Discours de haine | Le classificateur dit un
discours de haine] ?
 $p[C_H|C_H]$?

$$\frac{p[C_H|H] p[H]}{p[C_H]}$$

$$\frac{1 * 0.001}{p[C_H]}$$

Les maths:

Étant donné un texte, le classificateur dit "discours de haine".

Quelle est la probabilité qu'il le soit réellement ?

Law of total probability

$$P[C_H] = p[C_H|H]p(H) + p(C_H|B)p(B)$$

$$\frac{1 * 0.001}{1 * 0.001 + 0.05 + 0.999} = 2\%$$

Erreur du taux de base

Pr[Disc]

Notre classificateur :
Taux de vrais positifs de 100 %

Emporter

Le résultat du classificateur doit être
pris avec un grain de sel

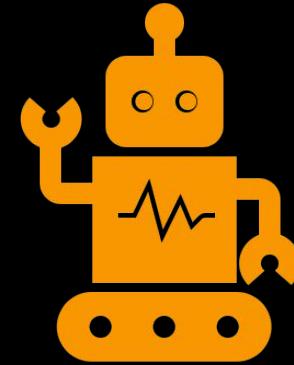
Les événements dont la fréquence est moyenne peuvent convenir, mais les événements rares (c'est-à-dire les signaux très faibles) sont difficiles à trouver

$p[\mathbb{C}_H]$

$$\frac{1 * 0.001}{1 * 0.001 + 0.05 + 0.999} = 2\%$$

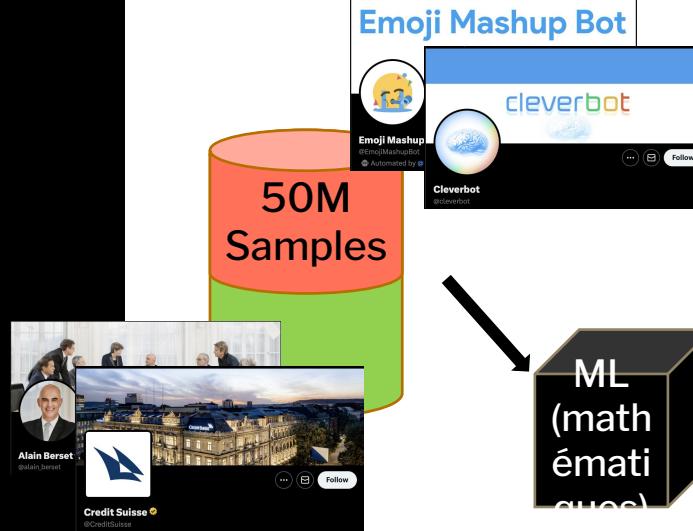
arqué comme
discours de haine
dit "discours
ellelement ?

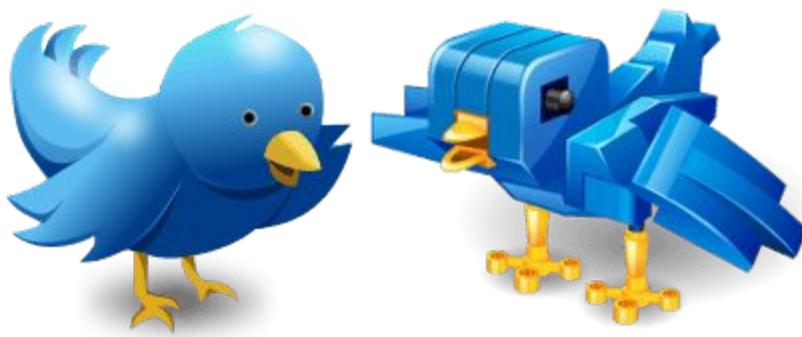
Détection de robot ?



Exemple d'apprentissage automatique : détection de robots

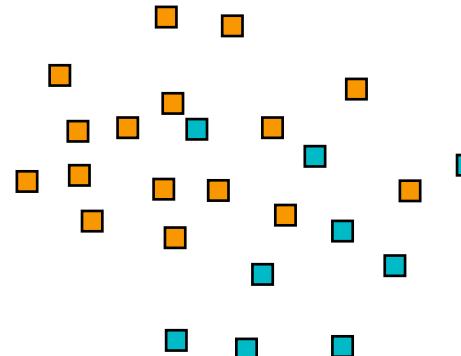
```
def detect_hate_speech_with_ML(text):
df = pd.read_csv('bot_examples.csv')
# transformer le texte en vecteur
vectoriseur = TfidfVectorizer()
data = vectorizer.fit_transform(df['text'])
labels = dataset['classe'].astype(int)
# apprentissage automatique
modele = RandomForest().fit(données,étiquettes)
    si model.predict(text):
return True # mot moyen trouvé ! :(
autre:
return False # aucun mot moyen n'a été trouvé :)
```





Répartition des erreurs

Se produit lorsque la plupart des erreurs du classificateur sont concentrées dans une sous-population / un groupe



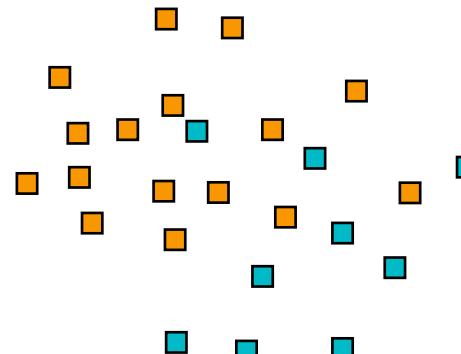
Répartition des erreurs

Se produit lorsque la plupart des erreurs du classificateur sont concentrées dans une sous-population/un groupe

Exemple : Détection de faux comptes Twitter

Orange = 😊

Bleu = 🤡



Répartition des erreurs

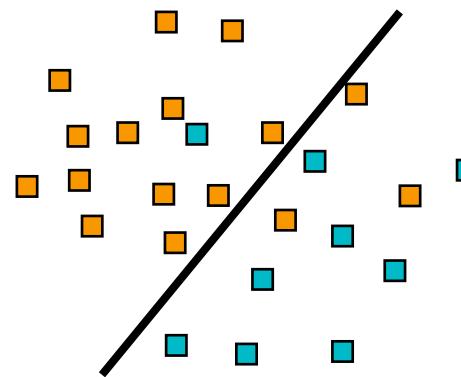
Se produit lorsque la plupart des erreurs du classificateur sont concentrées dans une sous-population/un groupe

Exemple : Détection de faux comptes Twitter

Orange = 😊

Bleu = 🤡

$(8+13)/25 = 84\%$ Précision



Répartition des erreurs

Se produit lorsque la plupart des erreurs du classificateur sont concentrées dans une sous-population/un groupe

Exemple : Détection de faux comptes Twitter

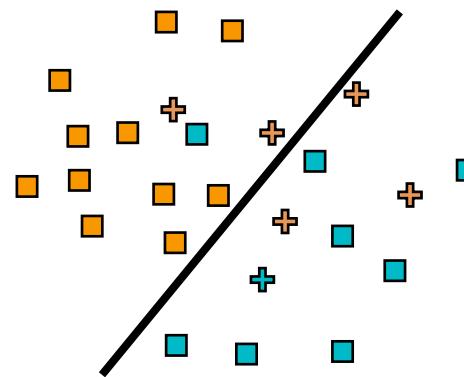
Orange = 😊

Bleu = 🤡

$(8+13)/25 = 84\%$ Précision

Croix = Politiciens (50%)

Carrés = Gens « normaux » (94 %)



26 Jan.

botometer.osome.iu.edu

Botometer

FAQ API Publications Bot Repo BEV Lite @FlorianGallwitz

Botometer®

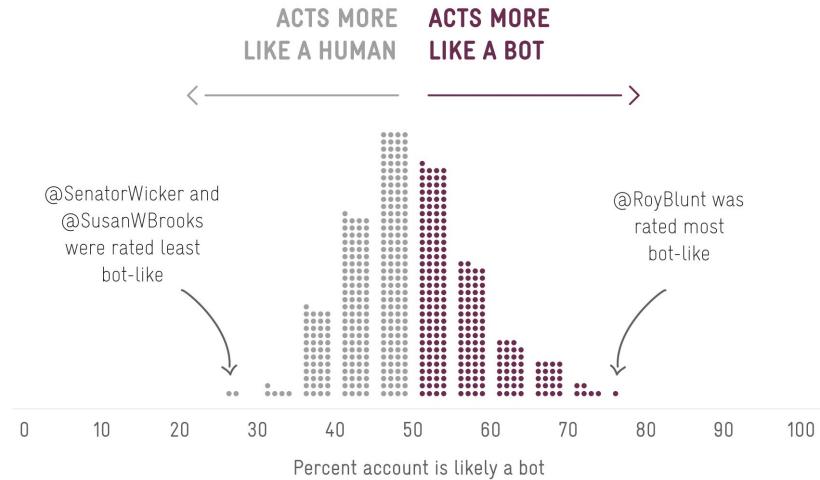
An OSoMe project (bot•o•meter)



Botometer (formerly BotOrNot) checks the activity of a Twitter account and gives it a score. Higher scores mean more bot-like activity. Use of this service requires Twitter authentication and permissions. ([Why?](#)) If something's not working or you have questions, please contact us only after reading the [FAQ](#). Botometer is a joint project of the Observatory on Social Media (OSoMe) and the Network Science Institute (IUNI) at Indiana University.

Check user Check followers Check friends

 @Reuters	3.6 / 5		X
 @POTUS	3.2 / 5		X
 @BernieSanders	3.4 / 5		X



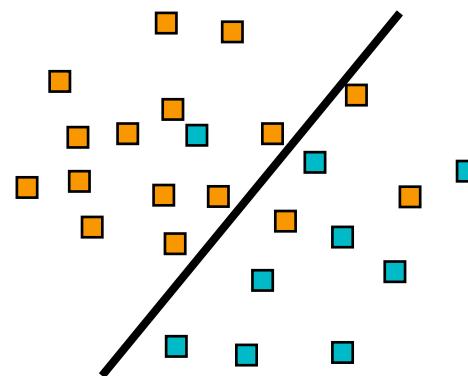
Décalage distributionnel

Se produit lorsqu'un classificateur est formé dans une zone et déployé dans une autre.

Exemple : Détection de faux comptes Twitter

Orange = 😊

Bleu = 🥶



Décalage distributionnel

Se produit lorsqu'un classificateur est formé dans une zone et déployé dans une autre.

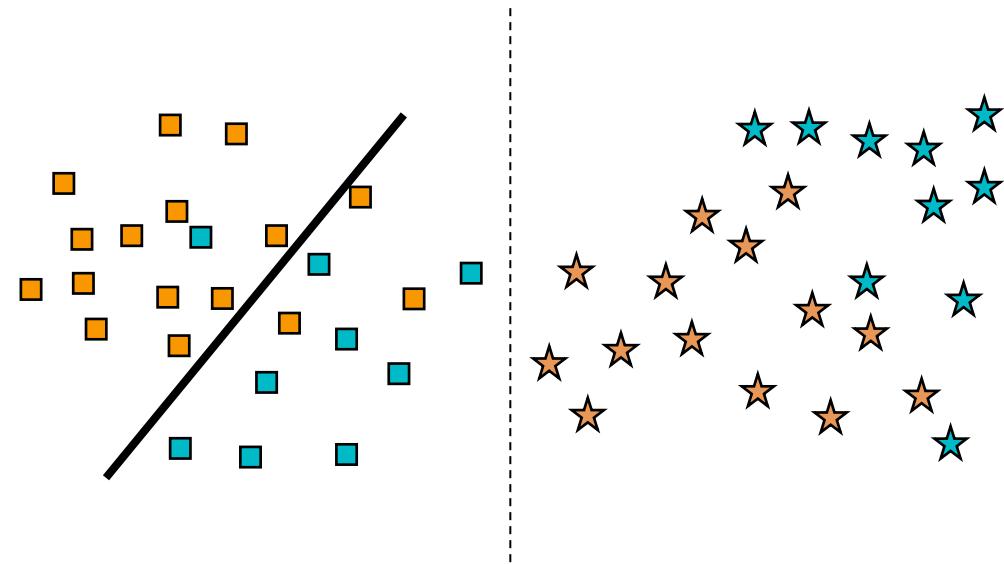
Exemple : Détection de faux comptes Twitter

Orange = 😊

Bleu = 🥶

Carrés = 🇺🇸

Étoiles = 🇷🇺



Décalage distributionnel

Se produit lorsqu'un classificateur est formé dans une zone et déployé dans une autre.

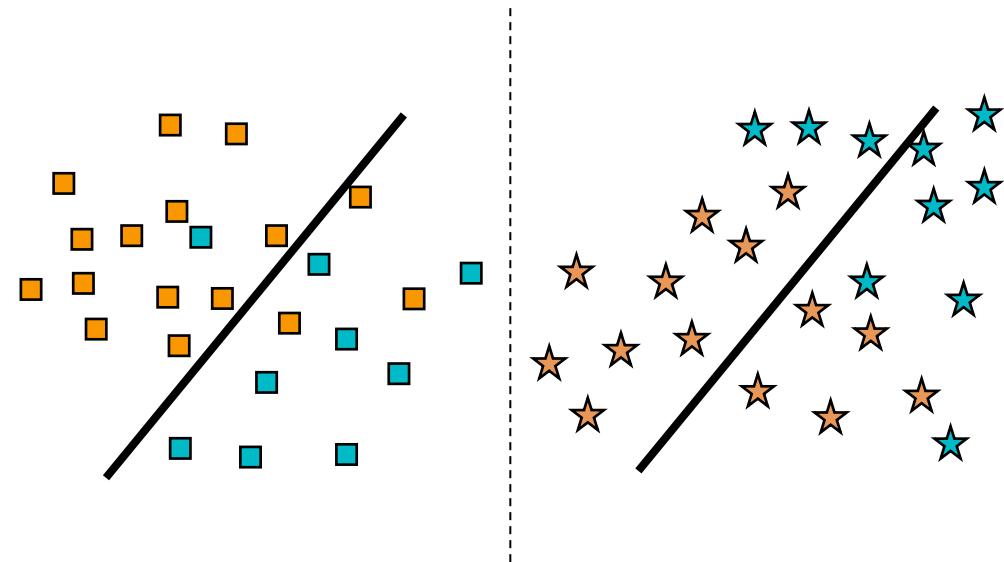
Exemple : Détection de faux comptes Twitter

Orange = 😊

Bleu = 🥶

Carrés = 🇺🇸

Étoiles = 🇷🇺



AVEC LE ML POUR LA DÉTECTI ON DES COMPOR TEMENT S ILLICITES (EN

Privacy

Bias

Distributional Shift

Distribute errors unfairly

Bias Reinforcement

Lack of transparency

Causation vs Correlation

Base Rate Fallacy

Adversarial datasets

Adversarial ML threats

Negative environmental outcomes

Antisocial outcomes

Externalize risks

Reward hacking

Misaligned performance metrics

Membership Inference

Lack of Explainability

...

Inflation du contenu

Manipuler la plate-forme pour rendre quelque chose plus populaire qu'il ne l'est réellement.

Inflation du contenu

Aimés de Facebook



Tendances Twitter



Votes pour Reddit



A large, stylized white bird logo, characteristic of the Twitter brand, is positioned on the left side of the slide. It is set against a solid blue background that occupies the left third of the frame.

Manipulation des tendances Twitter

Sujets tendance Twitter

Maskenpflicht

Top Latest People Photos Videos

p3likan @p3likaan · 41s
Mecklenburg-Vorpommern hat "klammheimlich" die **Maskenpflicht** im ÖPNV verschärft!
Bis jetzt waren medizinische Masken ausreichend.
Ab sofort nur noch ffp2!

msgiv.brandenburg.de
Landesregierung verlängert Corona-Verordnung
Die SARS-CoV-2-Infektionsschutz-Basismaßnahmenverordnung wird um...

Frettchen707 @frettchen707 · 46s
Wenn Lauterbach Verkehrsminister geworden wäre und seinen Jüngern empfohlen hätte Diesel statt Super zu tanken,ich glaube sehr viele hätten es gemacht 😂😂😂 #Maskenpflicht

Prof. Elon_Lab ✅ #ChecktheNarr... @Elon49943... · 47s
"Der Angriff des #Kartoffelvirus - Teil III"
Starring: Karl Lauterbach, DIVIGate, Deutsche Leidmedien
#Maskenpflicht

Bundesgesundheitsministerium ✅ @BMG_Bund · 11m

Search filters

People

- From anyone
 People you follow

Location

- Anywhere
 Near you

[Advanced search](#)

Trends for you

Trending
Impeachment
12.2K Tweets

Trending in Switzerland
Maskenpflicht
17.7K Tweets

Digital asset industry · Trending
#Ethereum
86.2K Tweets

Trending
#hacking
5,740 Tweets

Sujets tendance Twitter

"twitterragt açıldı"

Popüler **En Son** **Kişiler** **Fotoğraflar** **Videolar**

minnki @mimarinizgeldi · 39d
TwitterRaGt Açıldı merhabaaa 


mstfozgul @ozgulmst · 47d
TwitterRaGt Açıldı süper bir site arkadaşlar.twitter takipçilerinizi artırtın ak içün tıklayın ; twttgt.com


minnki @mimarinizgeldi · 1 sa
TwitterRaGt Açıldı ağalar bu ne en basa tutturulmus ama hiç flood yok!!


Arama filtreleri

Kişiler

Herhangi birinden

Takip ettiğin kişiler

Konum

Herhangi bir verde

Yakınlarında

Gelişmiş arama

Türkiye gündemleri

- 1 · Gündemdekiler
#FenerbahçeGeliyor
9.996 Tweet
 - 2 · Gündemdekiler
#benatatürküsevmiyorum
3.443 Tweet
 - 3 · Gündemdekiler
#salı
7.407 Tweet
 - 4 · Gündemdekiler
twitterragt açıldı

Sujets tendance Twitter

← "twitteragt açıldı"

Popüler En Son **Kişiler** Fotoğraflar Videolar

minnki @mimarinizgeldi · 39d Twitteragt Açıldı merhabaaa 🇺🇸🇹🇷

Quelque chose rend ces mots tendance.

Hypothèse : Il y avait des tweets qui contenaient ce mot, mais ils ont été supprimés

mstfozgul @ozgulmst · 47d Twitteragt açıldı süper bir site arkadaşlar.twitter takipçilerinizi artırmak için tıklayın ; tvigt.com

minnki @mimarinizgeldi · 1 sa Twitteragt açıldı agalar bu ne en basa tutturulmus ama hiç flood yok!!

Arama filtreleri

Kişiler

Herhangi birinden

Takip ettiğin kişiler

Konum

Herhangi bir yerde

Yakınlarında

Gelişmiş arama

Türkiye gündemleri

1 · Gündemdekiler

#FenerbahçeGeliyor

9.996 Tweet

2 · Gündemdekiler

#benatatürküsevmiyorum

3.443 Tweet

3 · Gündemdekiler

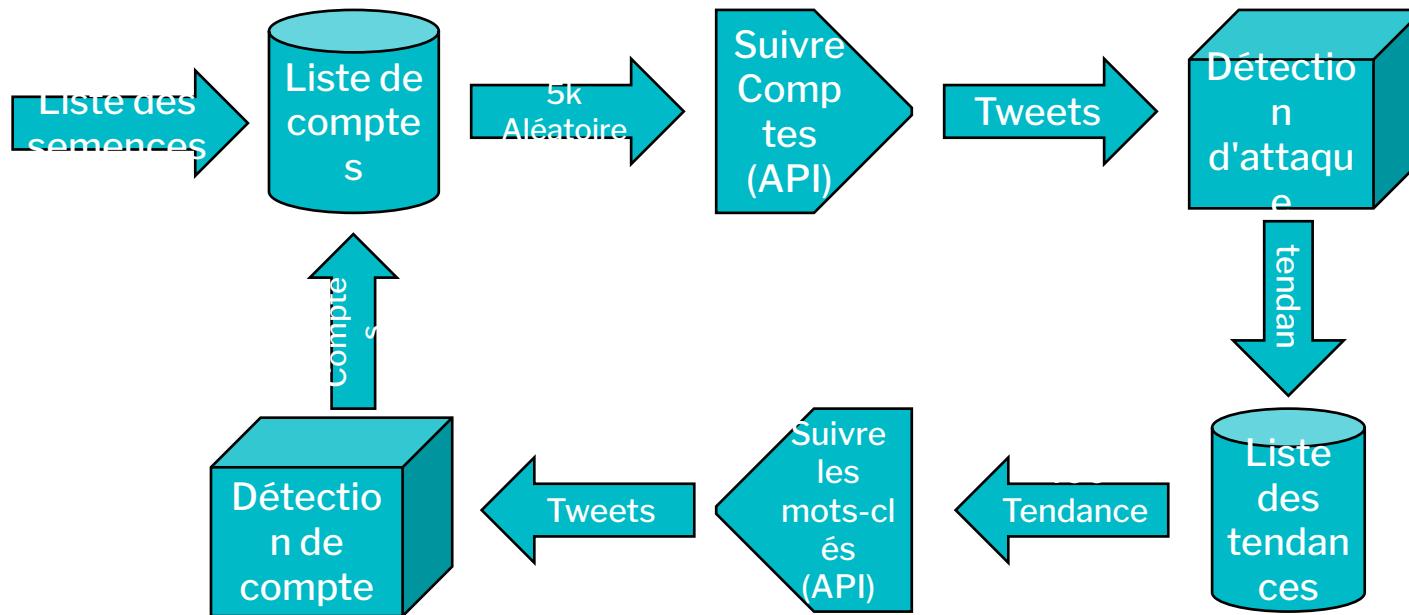
#sali

7.407 Tweet

4 · Gündemdekiler

twitteragt açıldı

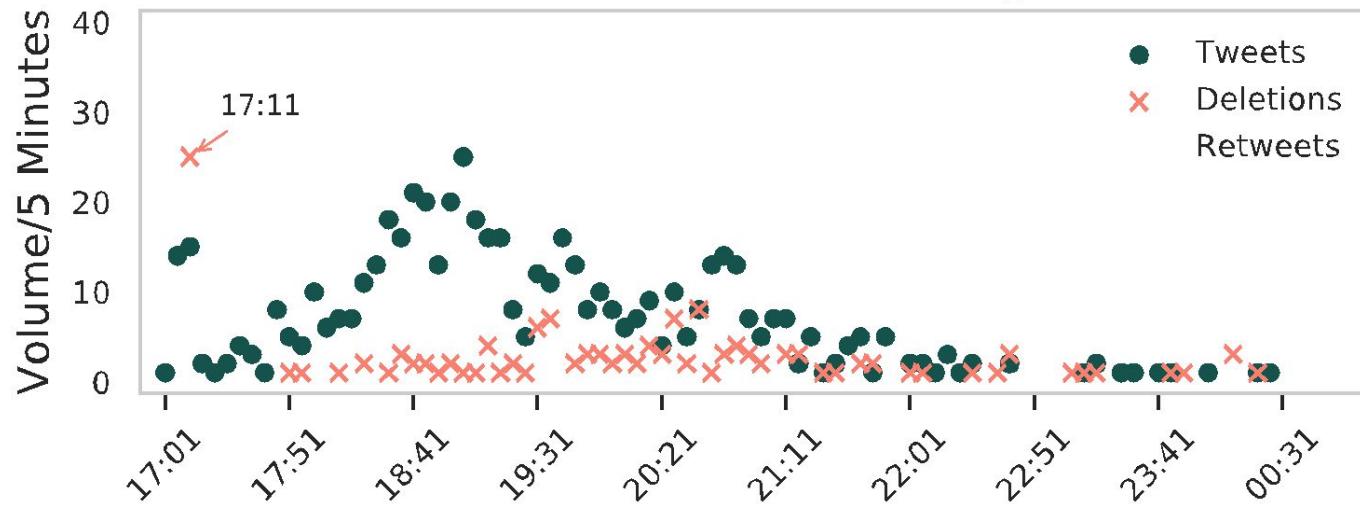
Collecte de données



Pot de miel

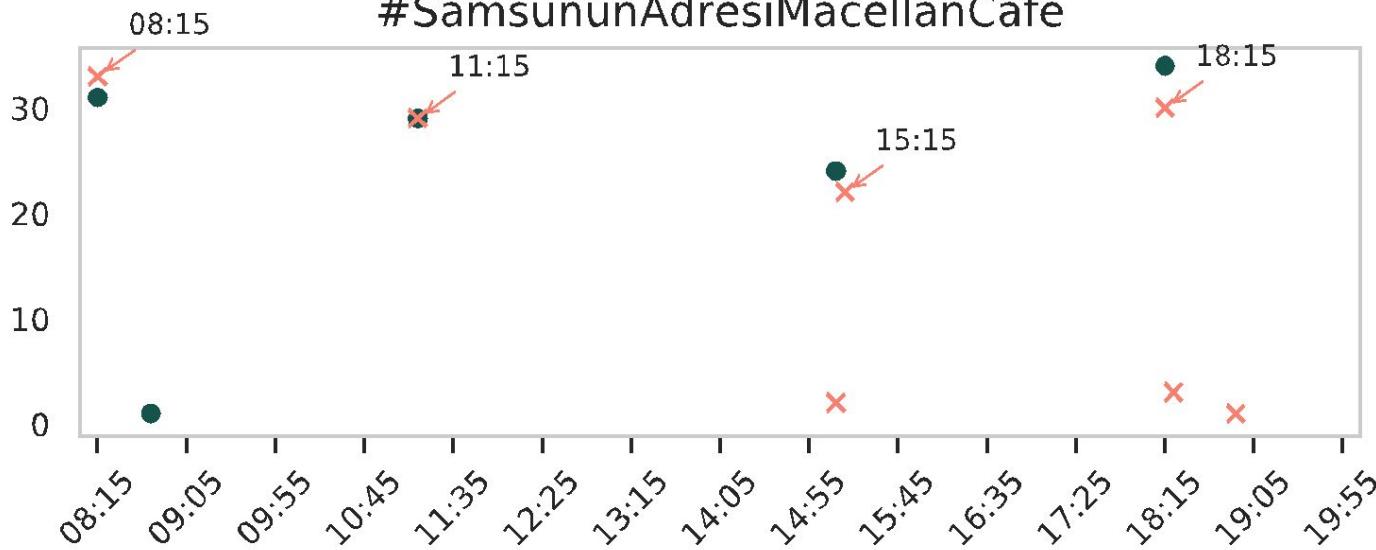


#İstanbulunUmudulmamoglu



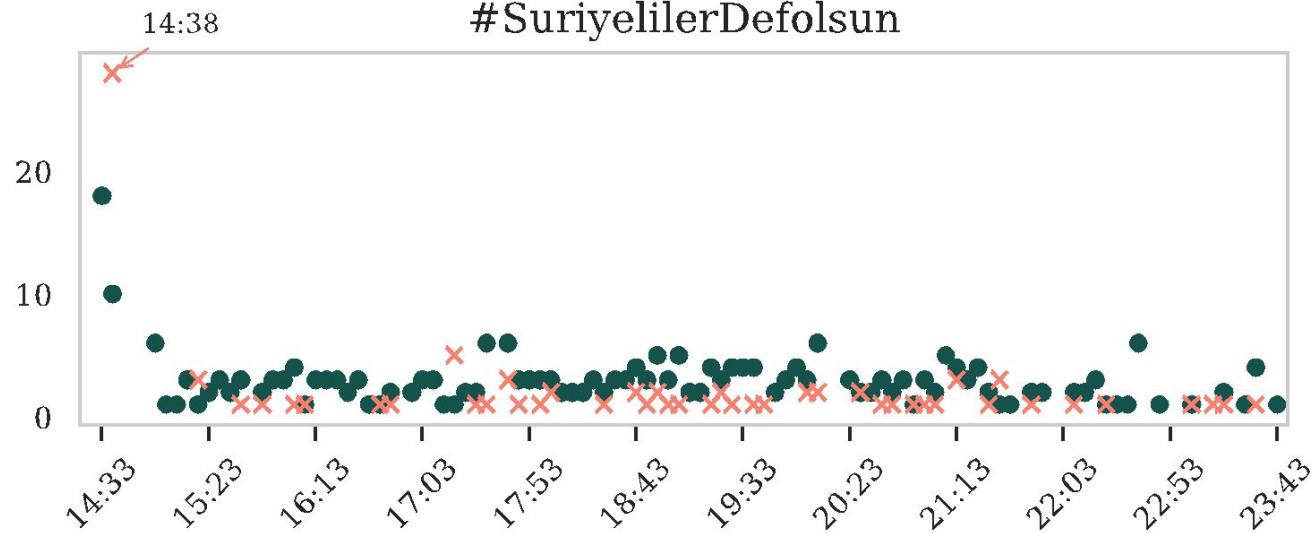
Exemples

#SamsununAdresiMacellanCafe



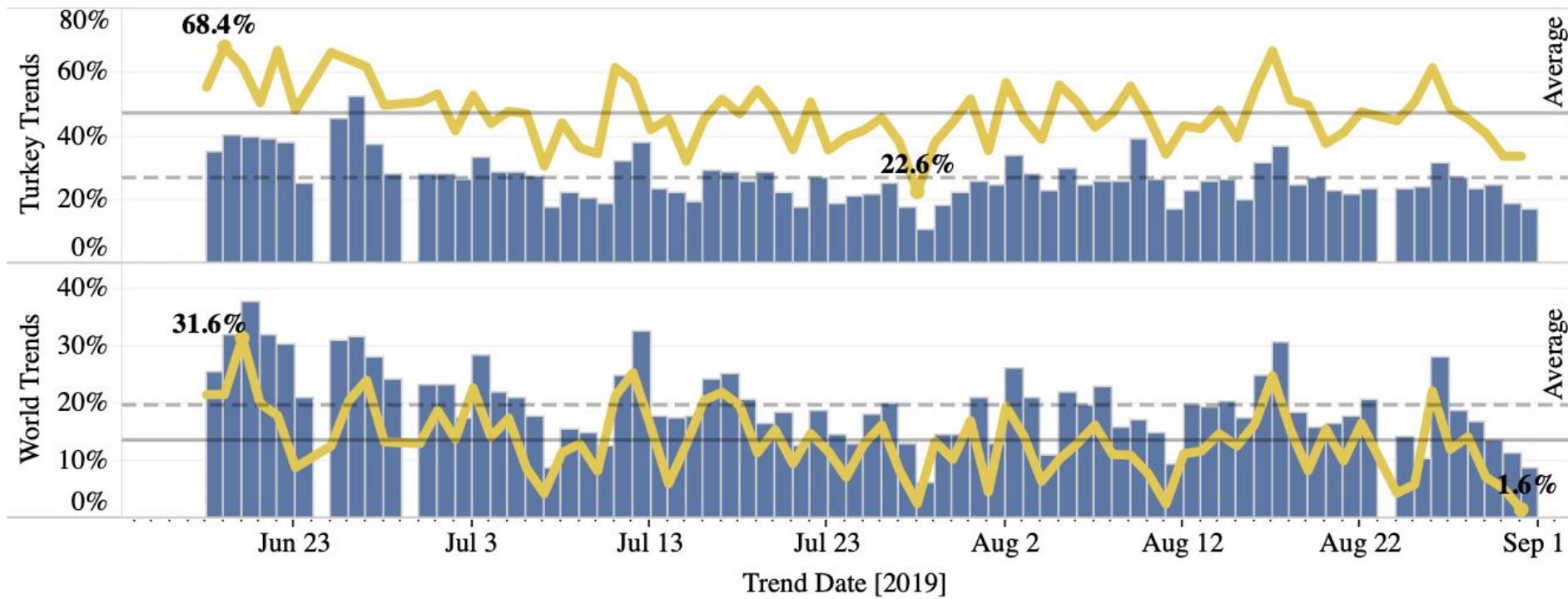
Exemples

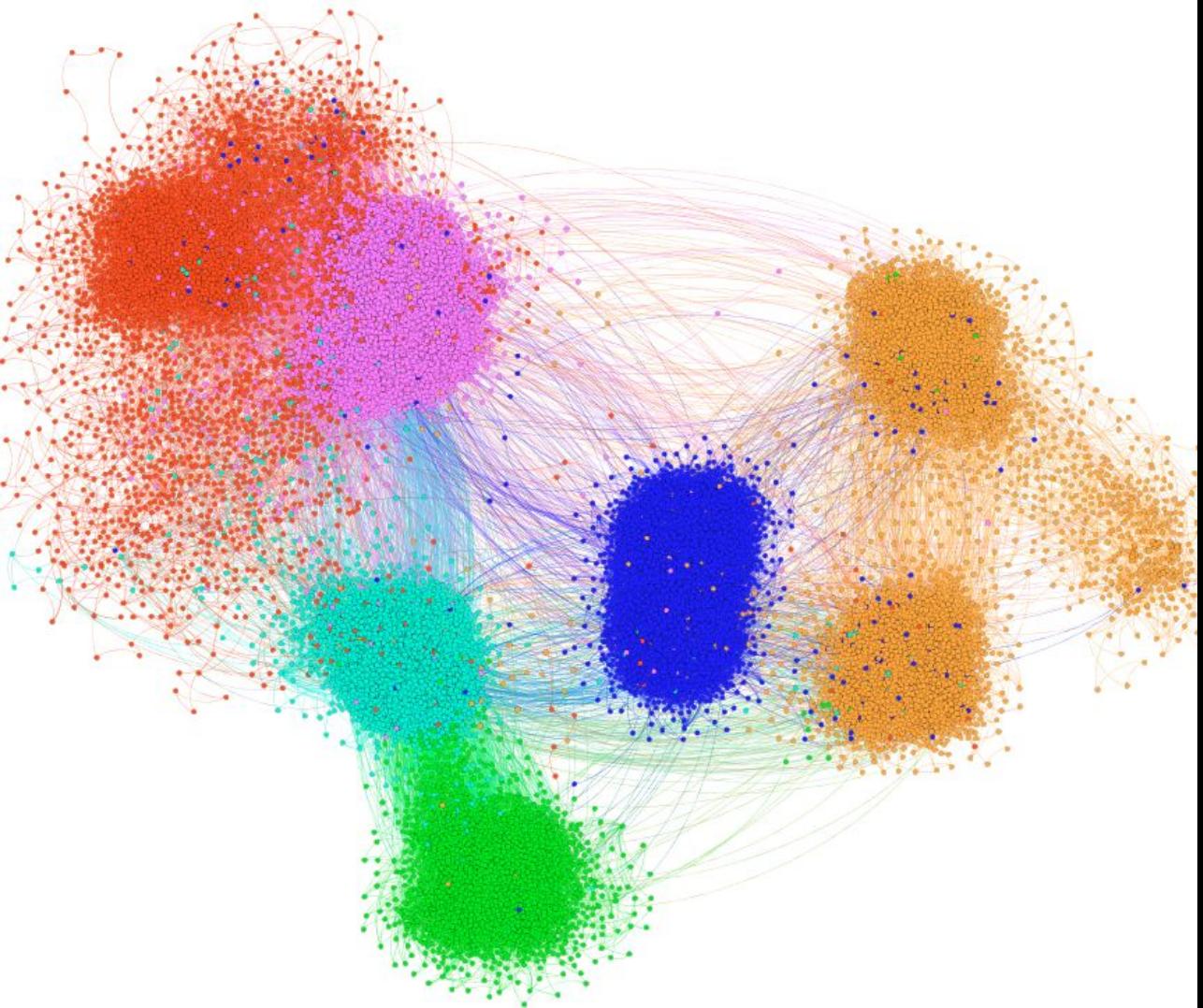
#SuriyelilerDefolsun



Exemples

Prévalence





Réseaux

Vert : Divers.
Cyan : Appel (Pardon)
Bleu : Publicités
(Spam)
Rose : Publicités
(Paris)
Rouge : Divers
Orange : Culte
(Furkan)



Manipulation
des likes sur
Facebook

HoneyPot

- 13 pages Facebook
- Ce n'est pas une vraie page, alors s'il vous plaît ne l'aimez pas.
- 5 ont été promus légitimement avec des publicités (États-Unis, France, Inde, Égypte et dans le monde)
- 8 pages ont été promues en utilisant 4 fermes similaires
- Budget : 6 \$/jour/compte



Que pouvons-nous apprendre de cela ?

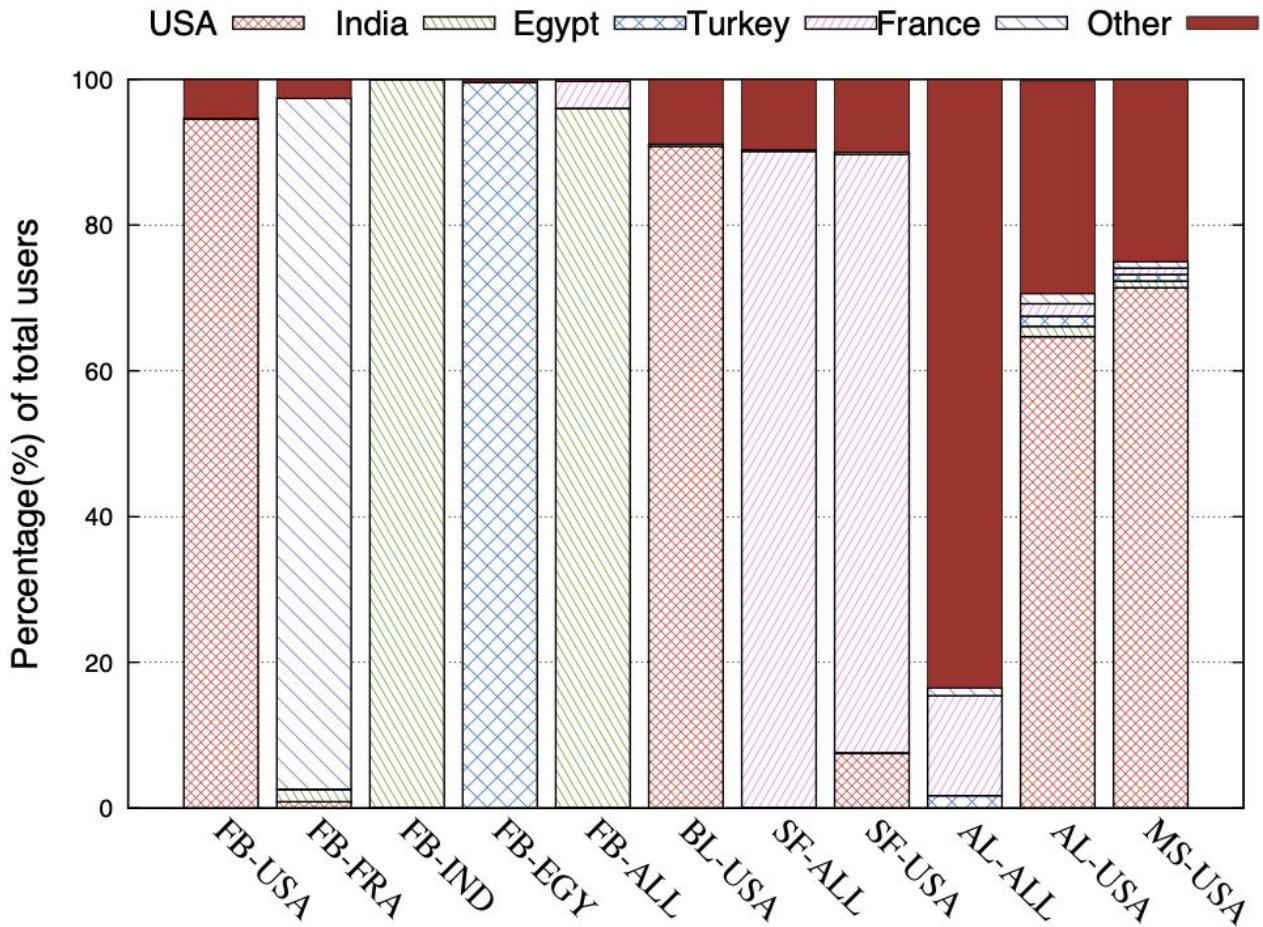
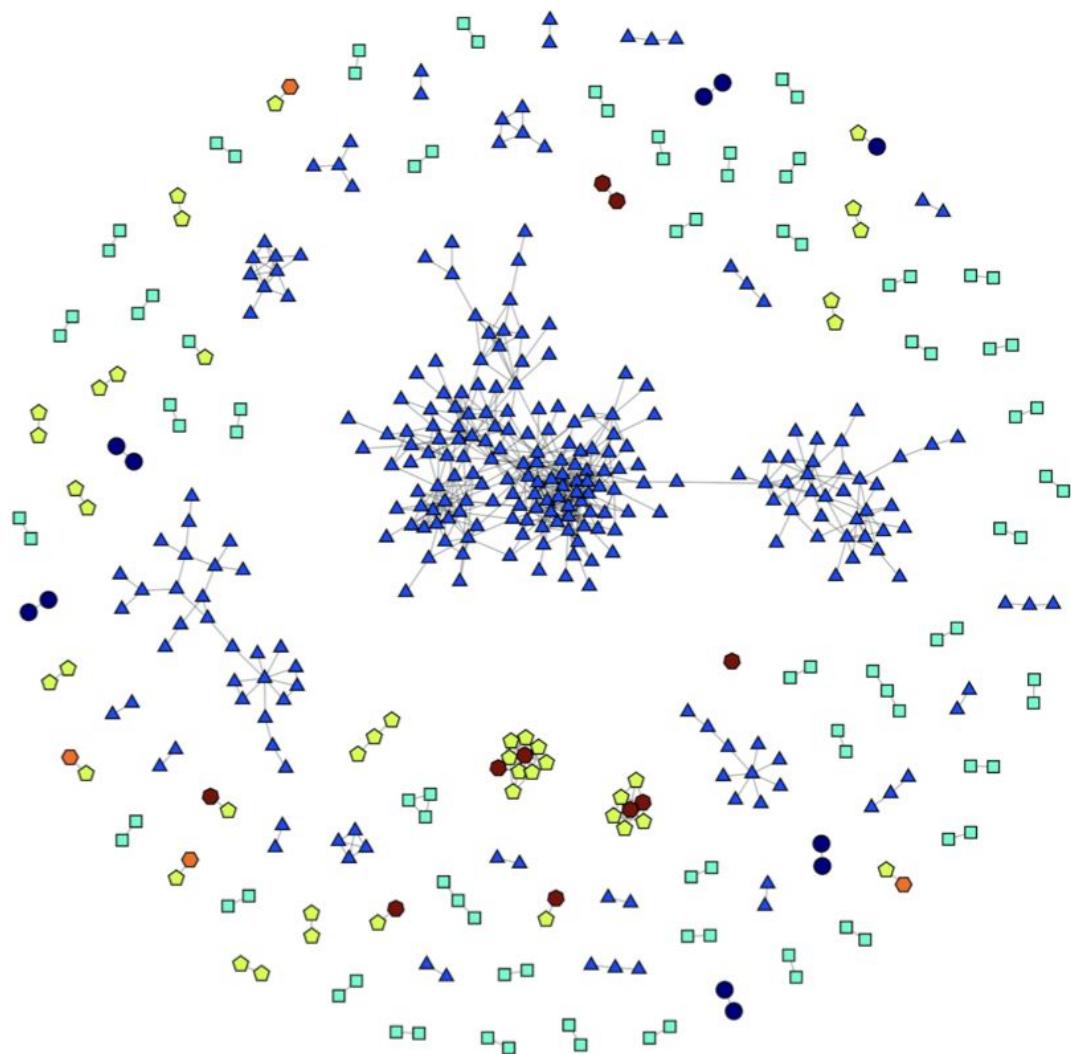


Figure 1: Geolocation of the likers (per campaign).

Que pouvons-nou s apprendre de cela ?

- Facebook
- ▲ BoostLikes
- SocialFormula
- ◆ AuthenticLikes
- MammothSocials
- ◆ AuthenticLikes +MammothSocial





Manipulation des
votes positifs de
Reddit

popular links

https://old.reddit.com/r/popular/?geo_filter=FR

MY SUBREDDITS - DASHBOARD - POPULAR - ALL - RANDOM - USERS - EDIT | add shortcuts from the my subreddits menu at left or click the button by the subreddit name, drag and drop to sort

Want to join? Log in or sign up in seconds. | English |

Reddit

POPULAR hot new rising controversial top gilded show images

Welcome to Reddit.
Come for the cats, stay for the empathy.

BECOME A REDDITOR and start exploring.

popular in: France

L éléphant dans la piece Écologie (self.france)
submitted 5 hours ago by Douillo0s to r/france
254 comments source share save hide report

PlayStation®5 jetzt verfügbar. Entdecke faszinierende Welten und erlebe fantastische Abenteuer. Beginne deine Reise noch heute! (playstation.com)
promoted by PlayStation Europe
1 promoted save report

Bonjour, j'ai reçu ce courrier, est-ce que c'est une arnaque ? Ask France (old.reddit.com)
submitted 2 hours ago by Marie_999 to r/france
93 comments share save hide report [I+c]

Pourquoi tant de personnes s'obstinent à boire de l'eau en bouteille chez elles alors que nous disposons de robinets et d'eau potable ? Ask France (i.redd.it)
submitted 13 hours ago by benzehdi to r/france
460 comments share save hide report [I+c]

Le bleu le moins bleu de ma vie Fait sans magouille avec mes couilles (i.redd.it)
submitted 6 hours ago by _Kantarasan_ to r/france
48 comments share save hide report [I+c]

Se faire insulter dans le métro pour avoir croisé ses jambes Culture (self.france)
submitted 18 hours ago by Spooktator to r/france
256 comments source share save hide report

La Société générale, BNP Paribas, Exane, Natixis et HSBC visées par des perquisitions dans un scandale de fraude fiscale hors norme Paywall (lemonde.fr)
submitted an hour ago by Ubik78 to r/france
22 comments share save hide report [I+c]

La finance internationale prend ses distances avec Macron - mediapart Paywall (mediapart.fr)
submitted an hour ago by tutatotu to r/france
32 comments share save hide report [I+c]

search

username password

remember me reset password login

Submit a new link

Submit a new text post

reddit premium

Get an ad-free experience with special benefits, and directly support Reddit.

Get Reddit Premium

Méthodologie : Simulation

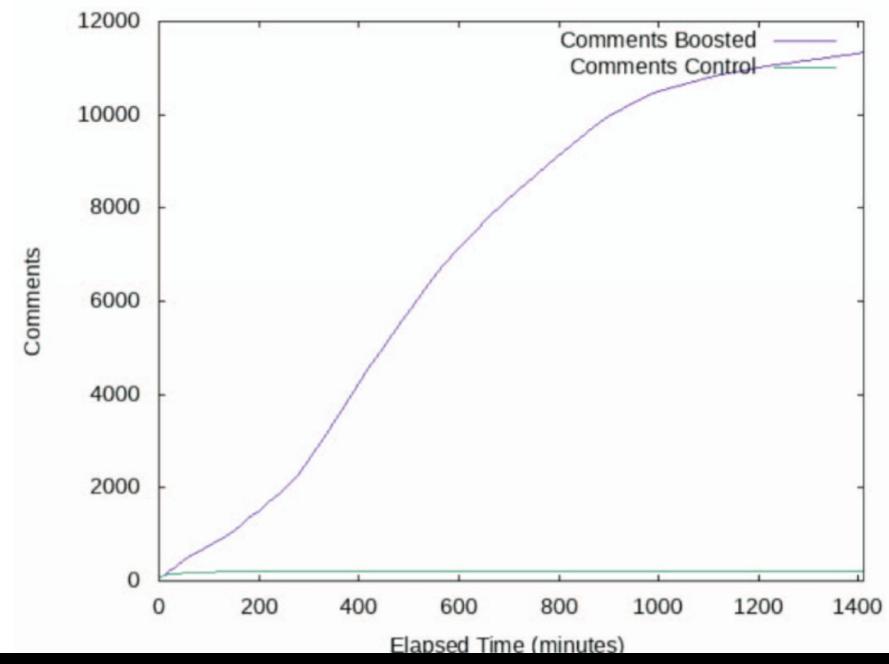
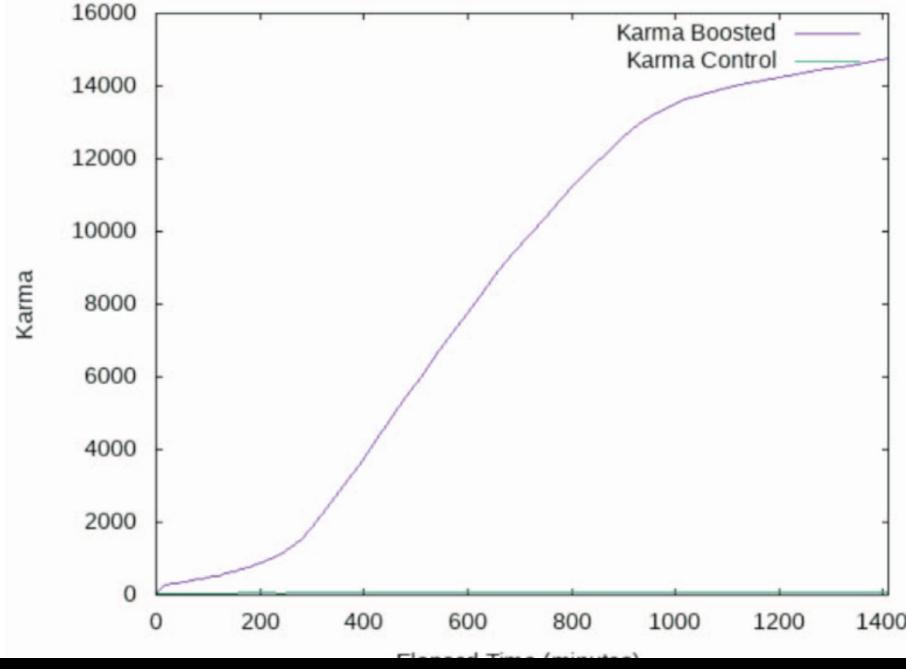
2 sous-reddits (r/the_donald et r/askreddit)

Tous les jours pendant 7 jours :

Les 50 messages les plus récents dans chaque subreddit

Suivre 25 postes de contrôle

Suivre 25 messages manipulés (ajouté 10 votes positifs au début)



Résultats

Présentation des méthodes

Apprentissage automatique

Centralité du graphe social

Modélisation de sujet

Détection communautaire

Récupération Web

Analyse des données

Méthodes mixtes (Qual)

Simulations de comportement illicite

Pots de miel

Collecte de données API

A emporter

Aucune méthode n'est unique

Nous avons besoin de méthodes spécialisées pour des problèmes uniques

L'apprentissage automatique peut être un outil précieux, mais il a ses inconvénients

Les graphiques sociaux sont parfaits pour étudier les communautés, dans le crime ou autre.

