Improving User Anonymity at the Content and Network Level

Sadia Afroz
University of California, Berkeley
How to be anonymous?
How to be anonymous?

I almost wish I hadn't gone down that rabbit-hole

blog.com
How to be anonymous?

Alice
alice@email.com

I almost wish I hadn't gone down that rabbit-hole
blog.com
How to be anonymous?

Alice
alice@email.com

I almost wish I hadn't gone down that rabbit-hole

blog.com
How to be anonymous?

I almost wish I hadn't gone down that rabbit-hole

Alice

alice@email.com

Alice’s IP
How to be anonymous?

Queen of hearts Alice

alice@email.com

I almost wish I hadn't gone down that rabbit-hole

blog.com
How to be anonymous?

Queen of hearts  Alice
xyz@email.com  alice@email.com

I almost wish I hadn't gone down that rabbit-hole
blog.com

Alice’s IP
How to be anonymous?

Queen of hearts  Alice

xyz@email.com  alice@email.com

Alice’s IP  “Random” IP  I almost wish I hadn't gone down that rabbit-hole blog.com
How to be anonymous?

Queen of hearts Alice

I almost wish I hadn't gone down that rabbit-hole

blog.com
How to be anonymous?

I almost wish I hadn't gone down that rabbit-hole

Alice

Queen of hearts

xyz@email.com

alice@email.com

5
I almost wish I hadn't gone down that rabbit-hole
I almost wish I hadn't gone down that rabbit-hole.
I almost wish I hadn't gone down that rabbit-hole
My Research

Adversary

Content anonymity
IEEE S&P '12,
IEEE S&P'14,
PETS '12

I almost wish I hadn't gone down that rabbit-hole

Adversarial learning
AISEC '12, 13, 14
TISSEC '12

blog.com
My Research

Practical privacy attacks
CCS '14

Adversarial learning
AISEC '12, 13, 14
TISSEC '12

Content anonymity
IEEE S&P '12,
IEEE S&P'14,
PETS '12

I almost wish I hadn't gone down that rabbit-hole

blog.com
My Research

Content anonymity
- IEEE S&P '12,
- IEEE S&P'14,
- PETS '12

Practical privacy attacks
- CCS '14

Adversarial learning
- AISEC '12, 13, 14
- TISSEC '12

I almost wish I hadn't gone down that rabbit-hole

Adversary
- Cybercrime, Politically motivated adversary
- ECRIME '13, FC '15
My Research

Adversary
Cybercrime, Politically motivated adversary

Practical privacy attacks
CCS '14

Content anonymity
IEEE S&P '12,
IEEE S&P'14,
PETS '12

Adversarial learning
AISEC '12, 13, 14
TISSEC '12

I almost wish I hadn't gone down that rabbit-hole
blog.com
Anonymity and writing style

• Can we link different identities using writing style?
Writing style analysis: Stylometry

Everybody’s writing style is unique
Writing style analysis: Stylometry

Everybody’s writing style is unique

Regional differences:
Couch vs Sofa
Writing style analysis: Stylometry

Everybody’s writing style is unique

Regional differences: Couch vs Sofa

Similar meaning but different words: Although vs Though
Current methods use machine learning

Cormac McCarthy

Ernest Hemingway
Current methods use machine learning

What's the bravest thing you ever did?

He spat in the road a bloody phlegm. Getting up this morning, he said.

Cormac McCarthy

Ernest Hemingway
What’s the bravest thing you ever did?

He spat in the road a bloody phlegm. Getting up this morning, he said.

He no longer dreamed of storms, nor of women, nor of great occurrences, nor of great fish, nor fights, nor contests of strength, nor of his wife.

Cormac McCarthy

Ernest Hemingway

Current methods use machine learning
Current methods use machine learning

What's the bravest thing you ever did?
He spat in the road a bloody phlegm. Getting up this morning, he said.

Cormac McCarthy

He no longer dreamed of storms, nor of women, nor of great occurrences, nor of great fish, nor fights, nor contests of strength, nor of his wife.

Ernest Hemingway
Current methods use machine learning

What's the bravest thing you ever did?

He spat in the road a bloody phlegm. Getting up this morning, he said.

Cormac McCarthy

Extract features

He no longer dreamed of storms, nor of women, nor of great occurrences, nor of great fish, nor of fights, nor of contests of strength, nor of his wife.

Ernest Hemingway

Extract features
Current methods use machine learning

What’s the bravest thing you ever did?

He spat in the road a bloody phlegm. Getting up this morning, he said.

He no longer dreamed of storms, nor of women, nor of great occurrences, nor of great fish, nor fights, nor contests of strength, nor of his wife.

Cormac McCarthy

Ernest Hemingway

Extract features

Freq of function words

Freq of punctuations
Current methods use machine learning

What's the bravest thing you ever did?

He spat in the road a bloody phlegm. Getting up this morning, he said.

He no longer dreamed of storms, nor of women, nor of great occurrences, nor of great fish, nor fights, nor contests of strength, nor of his wife.
Current methods use machine learning

Just remember that the things you put into your head are there forever, he said. You might want to think about that.
Current methods use machine learning

Just remember that the things you put into your head are there forever, he said. You might want to think about that.

Who wrote this? Model

Test document
Current methods use machine learning

Just remember that the things you put into your head are there forever, he said. You might want to think about that.

Who wrote this?

Model

Cormac McCarthy

Test document
Limitations of current methods

- Evaluated on formal English language text: books, emails (Abbasi et al., Koppel et al.) and blogs (Narayannan et al.)

- Known ground truth

- Not anonymity concerned
Underground Forums
Underground Forums

Understand the social networks of underground forums:

• Who is talking with whom?
• Who is trading with whom?
Underground Forums
Underground Forums

Buying malware
Underground Forums

Buying malware

Buying crypter
Underground Forums

Buying malware → Selling email/password

Buying crypter
Underground Forums

Buying malware

Selling email/password

Buying crypter
Underground Forums

- Buying malware
- Selling email/password
- Buying crypter
Underground Forums

- Buying malware
- Selling email/password
- Buying crypter
Underground forums

4 forums
Underground forums

4 leaked forums
Underground forums

4 leaked forums

- Antichat (Russian)
  - Password cracking
  - 41k users

- BlackhatWorld (English)
  - Blackhat seo
  - 8k users

- Carders, L33tCrew (German)
  - Stolen accounts
  - 8-12k users
Analyzing writing style is challenging!

- Not English, l33tsp3ak (e.g., L33t, pwn3d)

1337 down? **Neh, die Lösung!**
Ne klappt nit, denke mal eher das sie mal wieder DNS probleme haben
Analyzing writing style is challenging!

- Not only conversation

819 11/12 xxxx86xx Mxx Bxx mastercard
483 04/13 xxxx97xx Nxx Dxx mastercard
043 12/10 xxxx01xx Cxx Bxx mastercard
475 07/12 xxxx75xx Exx Sxx mastercard
Analyzing writing style is challenging!

- No ground truth
Doppelgänger Finder:
Cluster accounts that belong to a user
Doppelgänger Finder: Cluster accounts that belong to a user

Author A
Doppelgänger Finder: Cluster accounts that belong to a user

Author A

Author B
Doppelgänger Finder:
Cluster accounts that belong to a user

P(A wrote B’s doc)

Author A  →  Author B
Doppelgänger Finder: Cluster accounts that belong to a user

P(A wrote B’s doc) → P(B wrote A’s doc)
Doppelgänger Finder: Cluster accounts that belong to a user

Combined score, $T = P(A \text{ wrote } B\text{’s doc}) \times P(B \text{ wrote } A\text{’s doc})$

A and B are the same author if $T > \text{ threshold}$
Doppelgänger Finder: Cluster accounts that belong to a user

Author A

P(A wrote B’s doc) →
P(B wrote A’s doc)

Author B

Combined score, $T = P(A \text{ wrote } B\alpha) \times P(B \text{ wrote } A\alpha)$

A and B are the same author if $T >$ threshold
Doppelgänger Finder: Cluster accounts that belong to a user

Combined score, $T = P(\text{A wrote B's doc}) \times P(\text{B wrote A's doc})$

A and B are the same author if $T > \text{threshold}$

If A==B and B==C means A==B==C
Data with ground truth

- 100 English blogs
- Written by 50 authors, 2 blogs per author
- Collected by crawling Google+ profiles of the authors (Narayanan et al. 2012)
Probability scores of true pairs > false pairs
Probability scores of true pairs > false pairs

No true pair after this
Probability scores of true pairs > false pairs

Best threshold:
True pair = 48, False pair = 5

No true pair after this
Score on Carders

Combined Probability Scores vs. Rank of Author Pairs
Score on Carders

Our chosen threshold: 21 pairs
Our chosen threshold: 21 pairs
Manual analysis criteria

• To verify we looked at
  • Similar or not? : Username, ICQ, Signature, Contact information, Account information, Topics.
  • Do they talk with each other?
  • Others:
    • Do they acknowledge their other accounts?
    • Do they have common properties with some other users?
    • Were they banned from the forum?
Combined probability score on Carders

![Graph showing combined probability scores against rank of author pairs. The x-axis represents the rank of author pairs ranging from 0 to 100, and the y-axis represents combined probability scores ranging from 0.0 to 1.0. The data points are plotted as green dots, showing a decrease in probability scores as the rank increases.](image-url)
Combined probability score on Carders

Usernames: per**, Smi**

Acknowledge, same ICQ, sell weed
Combined probability score on Carders

Usernames: per**, Smi**

Acknowledge, same ICQ, sell weed

Usernames: Pri**, Lou**

Same ICQ, Topics
Combined probability score on Carders

Usernames: Kan**, deb**
Same ICQ

Usernames: per**, Smi**
Acknowledge, same ICQ, sell weed

Usernames: Pri**, Lou**
Same ICQ, Topics

Usernames: Kan**, deb**
Same ICQ
Combined probability score on Carders

Usernames: Kan**, deb**

Same ICQ

Usernames: per**, Smi**

Acknowledge, same ICQ, sell weed

Usernames: Pri**, Lou**

Same ICQ, Topics

Usernames: Kan**, deb**

Same ICQ

Usernames: Sch**, bob**
mfg, Kokain
Combined probability score on Carders

Usernames: Kan**, deb**
Acknowledge, same ICQ, sell weed

Usernames: per**, Smi**

Usernames: Pri**, Lou**
Same ICQ, Topics

Usernames: Kan**, deb**
Same ICQ

Usernames: Sch**, bob**
mfg, Kokain

Usernames: Mr.**, Fle**
Talk with each other

Combined Probability Scores

Rank of Author Pairs
Combined probability score on Carders

Usernames: per**, Smi**
Acknowledged, same ICQ, sell weed

Usernames: Kan**, deb**
Same ICQ

Usernames: Pri**, Lou**
Same ICQ, Topics

Usernames: Sch**, bob**
mfg, Kokain

Usernames: Mr.**, Fle**
Talk with each other

Usernames: puT**, pol**
Nothing matches

Combined Probability Scores

Rank of Author Pairs
We identified at least 10 pairs of doppelgängers. 5 more pairs without significant evidence.
Use of duplicate accounts

- Sockpuppet:
  - Raise fake demands for products
Use of duplicate accounts

• Sockpuppet:
  • Raise fake demands for products

I want to sell x
Use of duplicate accounts

• Sockpuppet:
  • Raise fake demands for products

I want to sell x
OMG!!!
I’ll buy them all
Use of duplicate accounts

• Accounts for sale:
  • Normal accounts: 10 €, 2nd level: 25 €, 3rd Level 50€
  • Moderators 100€ and admins 200€,
  • Whole database 500 €
The underground forums like our work!

We should run the following script on BHW
doppelganger-finder Doppelganger-finder finds multiple accounts (doppelgangers) of a user. It is useful to merge multiple aliases of a user.

https://github.com/sheetal57/doppelganger-finder

Summary: Doppelgänger Finder

• Future:
  • Clustering ads escort services (Darpa Memex)

• Code: https://github.com/sheetal57/doppelganger-finder

How to evade stylometry?

- Write less
- Write differently
- Change your style — Anonymouth
Evade stylometry: write differently

- We studied two ways to change writing style:
  - **Imitation**: imitate Cormac McCarthy
  - **Obfuscation**: writing in a different way
- We collected regular and deceptive documents using Amazon Mechanical Turk

Accuracy in regular documents is high
Accuracy in obfuscated writing

- 9-Feature (NN)
- Synonym-Based
- Writeprints Baseline (SVM)
- Random

Number of Authors

Accuracy values shown for different numbers of authors.
Accuracy in imitated writing

Number of Authors

- 9-Feature (NN)
- Synonym-Based
- Writeprints Baseline (SVM)
- Random
People can change their writing style!
People can change their writing style!

Maintaining it could be hard
A Gay Girl in Damascus

“Amina Arraf”
A Gay Girl in Damascus

Fake picture  
(copied from Facebook)

“Amina Arraf”
A Gay Girl in Damascus

Fake picture
(copied from Facebook)

The real “Amina”

“Amina Arraf”
A Gay Girl in Damascus

Fake picture (copied from Facebook)

“Amina Arraf”

The real “Amina”
A Gay Girl in Damascus

Fake picture (copied from Facebook)

“Amina Arraf”

The real “Amina”

Thomas MacMaster
A 40-year old American male
A Gay Girl in Damascus

I live in Damascus, Syria. It's a repressive police state. Most LGBT people are still deep in the closet or staying as invisible as possible. But I have set up a blog announcing my sexuality, with my name and my photo. Am I crazy? Maybe.
Authorship of the blog

Thomas MacMaster (as himself): 54%
Thomas MacMaster (as Amina Arraf): 43%
Other (Thomas’s wife, Britta): 3%

Anonymizing writing style

* Guide users to change their writing style
* Tell them how anonymous they are
Evade Stylometry: Change writing style

Anonymouth

https://psal.cs.drexel.edu/
How Anonymouth works

1. Background corpus
2. Document
How Anonymouth works

1. Background corpus
2. Document
   - Is the document anonymized? (Jstylo)
How Anonymouth works

1. Background corpus
2. Document
3. Is the document anonymized? (Jstylo)
   - Yes!
   - No
   - Which feature to change (Rank features)

Document
How Anonymouth works

1. Background corpus
2. Document
3. Is the document anonymized? (Jstyro)
4. Yes!
5. Document
6. No
7. Which feature to change (Rank features)
8. How to change feature (Feature clustering)
How Anonymouth works

1. Background corpus
2. Document
3. Is the document anonymized? (Jstylo)
4. Which feature to change (Rank features)
5. How to change feature (Feature clustering)

YES!

Document

Shows suggestion
How Anonymouth works

1. Background corpus
2. Document
3. Is the document anonymized? (Jstylo)
   - Yes! Document
   - No
4. Which feature to change (Rank features)
5. Shows suggestion
6. How to change feature (Feature clustering)
Happy families are all alike; every unhappy family is unhappy in its own way. Everything was in confusion in the Oblonskys’ house.
Example

Happy families are all alike; every unhappy family is unhappy in its own way. Everything was in confusion in the Oblonskys’ house.
Happy families are all alike; every unhappy family is unhappy in its own way. Everything was in confusion in the Oblonskys’ house.

Happy families are all alike. And, every family that isn’t happy, is unhappy in its own way. The Oblonskys’ house was in turmoil.
Anonymouth user study

• 10 participants
  – 6500+ pre-existing documents
  – 500-word document to modify
• Background corpus: 6 authors documents
• Classifier: 9-features and SVM
Does it work in theory?
Does it work in theory?

Document To Anonymized

Extract features

F0 F1 F2 F3 F4
Does it work in theory?

- Extract features: F0, F1, F2, F3, F4
- Rank features: F4, F3, F2, F0, F1

Document To Anonymized
Does it work in theory?

- Extract features: F0, F1, F2, F3, F4
- Rank features: F4, F3, F2, F0, F1
- Feature clustering: a, b, c, d, e
Does it work in theory?

Extract features → Rank features → Feature clustering

Document To Anonymized

F0 F1 F2 F3 F4

F4 F3 F2 F0 F1

a b c d e

Change features

`F4 `F3 `F2 `F0 `F1
Does it work in theory?

- Extract features: F0, F1, F2, F3, F4
- Rank features: F4, F3, F2, F0, F1
- Feature clustering: a, b, c, d, e
- Change features: 'F4, 'F3, 'F2, 'F0, 'F1
- Authorship confidence
Does it work in theory?

Extract features → Rank features → Feature clustering

Document To Anonymized

F0 F1 F2 F3 F4 → F4 F3 F2 F0 F1 → a b c d e

Change features

`F4 `F3 `F2 `F0 `F1

confidence > 0

Authorship confidence
In theory, Anonymouth works with any features.
In practice, Anonymouth works with 9-features
Summary: Anonymouth

- Code: https://psal.cs.drexel.edu/

- Paper:
  - [Best student paper] Use Fewer Instances of the Letter “i”: Toward Writing Style Anonymization. PETS 2012
  - [PET award 2013] Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity. TISSEC '12
Website Fingerprinting Attack
Website Fingerprinting Attack
Website Fingerprinting Attack

Where is Alice going?
Website Fingerprinting Attack

Where is Alice going?
Website Fingerprinting Attack

Where is Alice going?
Website Fingerprinting Attack

Where is Alice going?
Website Fingerprinting Attack

Where is Alice going?
How does it work?
How does it work?
How does it work?
How does it work?
How does it work?
How does it work?
How does it work?

Extract features

Extract features
How does it work?
How does it work?
How does it work?
How does it work?

Some page
How does it work?

Some page

What is this page?

Model
How does it work?

Some page

What is this page?

Model

64
Why is WF so important?

- Tor as the most advanced anonymity network
- Allows an adversary to discover the browsing history
- Series of successful attacks
- Low cost to the adversary

Number of top conference publications on WF (25)

![Number of top conference publications on WF (25)]
How practical is this attack?
How practical is this attack?
How practical is this attack?

Visit sites
How practical is this attack?

Visit sites → Collect packets
How practical is this attack?

Visit sites -> Collect packets -> Train model
How practical is this attack?

- Visit sites
- Collect packets
- Train model
- Test model
Unrealistic assumptions
Unrealistic assumptions

Client settings: e.g., browsing behavior
Website fingerprinting attack with multi-tab browsing

77.08%
Website fingerprinting attack with multi-tab browsing

Control:

- Test (0.5s): 9.8%
- Test (3s): 7.9%
- Test (5s): 8.23%

Overall: 77.08%
Website fingerprinting attack with multi-tab browsing

Control

Test (0.5s)

Test (3s)

Test (5s)

77.08%

9.8%

7.9%

8.23%

Tab 1

Tab 2

Time

BW
Unrealistic assumptions
Unrealistic assumptions
Unrealistic assumptions

Adversary: e.g., replicability

Website
Website fingerprinting attack with different Tor versions

- Coexisting Tor Browser Bundle (TBB) versions
- Versions: 2.4.7, 3.5 and 3.5.2.1 (changes in RP, etc.)
Website fingerprinting attack with different Tor versions

- Coexisting Tor Browser Bundle (TBB) versions
- Versions: 2.4.7, 3.5 and 3.5.2.1 (changes in RP, etc.)

Control (3.5.2.1)
Website fingerprinting attack with different Tor versions

- Coexisting Tor Browser Bundle (TBB) versions
- Versions: 2.4.7, 3.5 and 3.5.2.1 (changes in RP, etc.)
Unrealistic assumptions

Web: e.g., staleness

Website
Website staleness

Accuracy (%) vs Time (days)
Website staleness

Accuracy (%) vs. Time (days)

Less than 50% after 9d.
Theoretical Accuracy vs. Practical Accuracy

A Critical Evaluation of Website Fingerprinting Attacks. **ACM CCS 2014**
Future Work

• Machine Learning in Security:
  • Malware detection
  • Different metric for success: Adversarial robustness
Future Work

- Censorship circumvention
  - Evaluation of censorship circumvention tools
  - Tor as second class citizen
My Research

System

Adversary

Anonymity
IEEE S&P '12, IEEE S&P'14, PETS '12

Practical privacy attacks
CCS '14

Adversarial learning
AISEC ’12, 13, 14
TISSEC '12

Cybercrime, Politically motivated Adversary
ECRIME ’13, FC015