Stylometry and Online Underground Markets

Sadia Afroz, Aylin Caliskan Islam
Co-authors: Ariel Stolerman, Rachel Greenstadt, Damon McCoy
Previously at CCC...

• 26c3:
  • Introduced the concept of adversarial stylometry:
  • Authorship recognition algorithms can be evaded by changing writing style.
Previously at CCC...

- **26c3:**
  - Introduced the idea of adversarial stylometry
  - Authorship recognition algorithms can be evaded by changing writing style.

- **28c3:**
  - Released two tools:
    - JStylo (authorship recognition tool) and
    - Anonymouth (authorship anonymization tool)
This talk

• How stylometric analysis can be used in real world datasets?
  • Identify people based on writing style
  • Identify topic of their discussion
• Real world dataset: online underground markets
Overview

• Online underground markets
• Analysis
• Limitation and Challenges
• Future work
• Anonymouth
Online Underground Markets

- Underground market is a trading place to trade various stolen goods and/or tools such as exploits, malware repackaging kits, phishing kits.
Millions of LinkedIn passwords reportedly leaked online

A hacker says he's posted 6.5 million LinkedIn passwords on the Web -- hot on the heels of security researchers' warnings about privacy issues with LinkedIn's iOS app.

**Update 1:08 p.m. PT: LinkedIn confirms that passwords were "compromised."

LinkedIn users could be facing yet another security problem.

A user in a Russian forum says that he has hacked and uploaded almost 6.5 million LinkedIn passwords, according to The Verge. Though his claim has yet to be confirmed, Twitter users are already reporting that they've found their hashed LinkedIn passwords on the list, security expert Per Thorsheim said.

LinkedIn revealed through its own tweet that it's looking into reports of stolen passwords, and it advised users to stay tuned for more information.
Millions of LinkedIn passwords reportedly leaked online

A hacker says he's posted 6.5 million LinkedIn passwords on the Web -- hot on the heels of security researchers' warnings about privacy issues with LinkedIn's iOS app.

Update 1:08 p.m. PT: LinkedIn confirms that passwords were "compromised."

LinkedIn users could be facing yet another security problem.

A user in a Russian forum says that he has hacked and uploaded almost 6.5 million LinkedIn passwords, according to The Verge. Though his claim has yet to be confirmed, Twitter users are already reporting that they've found their hashed LinkedIn passwords on the list, security expert Per Thorsheim said.

LinkedIn revealed through its own tweet that it's looking into reports of stolen passwords, and it advised users to stay tuned for more information.
Why is this interesting?

- Cyber-underground ecosystem
- Key information about who controls a given bot
- Who maintains certain tools
- Size and scope of these markets.
A Closer Look at Two Bigtime Botmasters

Over the past 18 months, I've published a series of posts that provide clues about the possible real-life identities of the men responsible for building some of the largest and most disruptive spam botnets on the planet. I've since done a bit more digging into the backgrounds of the individuals thought to be responsible for the Rustock and Waledac spam botnets, which has produced some additional fascinating and corroborating details about these two characters.

In March 2011, KrebsOnSecurity featured never-before-published details about the financial accounts and nicknames used by the Rustock botmaster. That story was based on information leaked from SpamIt, a cybercrime business that paid spammers to promote rogue Internet pharmacies (think Viagra spam). In a follow-up post, I wrote that the Rustock botmaster's personal email account was tied to a domain name german.ru, which at one time featured a résumé of a young man named Dmitri A. Sergeev.

11 A Closer Look at Two Bigtime Botmasters

Over the past 18 months, I've published a series of posts that provide clues about the possible real-life identities of the men responsible for building some of the largest and most disruptive spam botnets on the planet. I've since done a bit more digging into the backgrounds of the individuals thought to be responsible for the Rustock and Waledac spam botnets, which has produced some additional fascinating and corroborating details about these two characters.

In March 2011, KrebsOnSecurity featured never-before-published details about the financial accounts and nicknames used by the Rustock botmaster. That story was based on information leaked from SpamIt, a cybercrime business that paid spammers to promote rogue Internet pharmacies (think Viagra spam). In a follow-up post, I wrote that the Rustock botmaster's personal email account was tied to a domain name germes.ru, which at one time featured a résumé of a young man named Dmitri A. Sergeev.
Online Underground Markets

• Two main markets:
  • Internet Relay Chat (IRC)
  • Web forums
Online Underground Markets

• Two main markets:
  • Internet Relay Chat (IRC)
  • Web forums
Forums

- Antichat - Russian forum (May 2002-Jun 2010)
- BadHacke - English/Hindi (Nov 2003-May 2008)
- Carders - German (Feb 2009- Dec 2010)
- L33tCrew - German (May 2007-Nov 2009)
How did we get the data?

- Leaked by anonymous people
- Publicly available
Members

<table>
<thead>
<tr>
<th>Community</th>
<th>Active members</th>
<th>Lurkers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antichat</td>
<td>25,871</td>
<td>15,165</td>
</tr>
<tr>
<td>Badhacke</td>
<td>5,123</td>
<td>9,306</td>
</tr>
<tr>
<td>Blackhat</td>
<td>4,489</td>
<td>3,097</td>
</tr>
<tr>
<td>Carders</td>
<td>4,229</td>
<td>3,026</td>
</tr>
<tr>
<td>L33tCrew</td>
<td>5,328</td>
<td>9,528</td>
</tr>
</tbody>
</table>
How Transaction Happens

i am selling my large targetted email list of more than 22.5 million emails
Challenges

41,036 Users

2,160,815 Public posts

194,498 Private msgs
Challenges

This is just one forum!

41,036 Users

2,160,815 Public posts

194,498 Private msgs

Friday, December 28, 12
Challenges

интересует взлом

41,036 Users

2,160,815 Public posts

194,498 Private msgs
Challenges

интересует взлом

Alles läuft vor eurem PC ab

41,036 Users

2,160,815 Public posts

194,498 Private msgs
Challenges

интересует взлом

I gave 40 bucks and no program what is up man?

Alles läuft vor eurem PC ab

2,160,815 Public posts

194,498 Private msgs

41,036 Users
Analysis

- Interaction network analysis
- Member profiling using writing style
- Topic discovery
Analysis

- Interaction network analysis
- Member profiling using writing style
- Topic discovery
Interaction Network Analysis

Represent a forum with a graph, $G=(V, E)$ where each user is a vertex.
Represent a forum with a graph, $G=(V, E)$ where each user is a vertex and each private message is an edge.
Interaction Network Analysis

• Goal:
  • Structure of interaction
  • Identify central members
Interaction Network Analysis

• Goal:
  • Structure of interaction
  • Identify central members:
    • Eigenvector centrality:
      • It is a measure of the influence of a node in a network.
      • Higher score == More influential member
Interaction Network Analysis: Antichat

With all users
Interaction Network Analysis: Antichat

With all users

Darker nodes send/receive more msgs

Bigger nodes are more influential

With all users
Interaction Network Analysis: Antichat

With all users

With top 50 influential users
Interaction Network Analysis: Badhacke

With all members  With top influential members
Analysis

- Interaction network analysis
- Member profiling using writing style
- Topic discovery
Profiling using writing style

• Everybody has a *unique* writing style.

• Goal:
  • Identify members using their writing style
Try JStylo

Accuracy in detecting authorship of regular documents

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

* Thanks to Michael Brennan for the graph

Friday, December 28, 12
Profiling using writing style

- Combine all messages of a member
- Find members with enough texts

JStylo

- Extract features
- Train classifier to model a member
- Evaluate
How much text is enough?
How much text is enough?: at least **5000** words

![Graph showing accuracy over the number of 500-word documents](image)

- **Carders (private)**
- **Carders (public)**
- **L33tcrew (private)**
- **L33tcrew (public)**
- **Antichat**
- **Blackhat**

Friday, December 28, 12
How much text is **enough?**
we used **6500 words**

---

**Accuracy**

**Number of 500-word documents**

- Carders (private)
- Carders (public)
- L33tcrew (private)
- L33tcrew (public)
- Antichat
- Blackhat

Friday, December 28, 12
Profiling using writing style

Combine all messages of a member → Find members with enough texts

JStylo

Extract features → Train classifier to model a member → Evaluate
What are these features?

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

Example from Carders
What are these features?

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

Example from Carders

*Translation

1337 down? Neh **, the solution! **
Ne nit works, rather guess they have again DNS problems

*Using Google translator*
What are these features?

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

Example from Carders

Freq. of n-grams
What are these features?

Example from Carders

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

Freq. of n-grams
Freq. of punctuations
What are these features?

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

Example from Carders

Freq. of n-grams
Freq. of punctuations
Freq. of special characters
What are these features?

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

Example from Carders

Freq. of n-grams
Freq. of punctuations
Freq. of special characters
Language Independent
What are these features?

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

Example from Carders

Language specific

Parts of speech
Freq. Function words
Freq. of ngrams
Freq. of punctuations
Freq. of special characters
Not all conversation

Bankname: XX
CCNumber: XXXXXXXX
CCHolder: XX XXXX
CCExpire: X / XXXX
CVV2: XX
Vorname: XX
Nachame: YY
Adresse: XXXXX
Stadt: XXXX
PLZ: XXXX
Land: XX
Telefon: XXXXX-XXXXX
E-mail: [email]victim@example.com[/email]
Geburtsdatum: XX / XX / XXXX
What’s wrong with that?

Member X

Bankname: AA
CCNumber: XXXXXXXX

Member Y

Bankname: AA
CCNumber: XXXXXXXX
What’s wrong with that?

Bankname: AA
CCNumber: XXXXXXXX

Bankname: AA
CCNumber: XXXXXXXX

Member X

Member Y

Both will look the same to a linguistic classifier
Profiling using writing style

Combine all messages of a member → Identify Product → Find members with enough texts

JStylo

Extract features → Train classifier to model a member → Evaluate
Identify Products

1. Product information has repeated patterns
2. Conversation usually has verb

Bankname: XX
CCNumber: XXXXXXXX
CCHolder: XX XXXX
CCExpire: X / XXXX
CVV2: XX
Vorname: XX
Nachname: YY
Address: XXXXX
Stadt: XXXX
PLZ: XXXX
Land: XX
Telefon: XXXXX-XXXXX
E-mail: [email]victim@example.com[/email]
Geburtsdatum: XX / XX / XXXX

Product

Conversation

1337 down? **Neh, die Lösung!
**
Ne klappt nit, denke mal eher
das sie mal wieder DNS
probleme haben
Identify Products

1. Product information has repeated patterns
2. Conversation usually has verb

<table>
<thead>
<tr>
<th>Bankname: XX</th>
<th>1337 down? <strong>Neh, die Lösung!</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>CCNumber: XXXXXXXX</td>
<td>**Ne klappt nit, denke mal eher</td>
</tr>
<tr>
<td>CCHolder: XX XXXX</td>
<td>das sie mal wieder DNS probleme</td>
</tr>
<tr>
<td>CCExpire: X / XXXX</td>
<td>habn**</td>
</tr>
<tr>
<td>CVV2: XX</td>
<td></td>
</tr>
<tr>
<td>Vorname: XX</td>
<td></td>
</tr>
<tr>
<td>Nachname: YY</td>
<td></td>
</tr>
<tr>
<td>Addresse: XXXXX</td>
<td></td>
</tr>
<tr>
<td>Stadt: XXXX</td>
<td></td>
</tr>
<tr>
<td>PLZ: XXXX</td>
<td></td>
</tr>
<tr>
<td>Land: XX</td>
<td></td>
</tr>
<tr>
<td>Telefon: XXXXX-XXXXX</td>
<td></td>
</tr>
<tr>
<td>E-mail: [email]<a href="mailto:victim@example.com">victim@example.com</a>[/email]</td>
<td></td>
</tr>
<tr>
<td>Geburtsdatum: XX / XX / XXXX</td>
<td></td>
</tr>
</tbody>
</table>

Product

Conversation
Identify Products

Bankname: XX
CCNumber: Xxxxxxxxx
CCHolder: XX XXXX
CCExpire: X / XXXX
CVV2: XX
Vorname: XX
Nachname: YY
Adresse: XXXX
Stadt: XXXX
PLZ: XXXX
Land: XX
Telefon: XXXX-XXXXX
E-mail: [email]victim@example.com[/email]
Geburtsdatum: XX / XX / XXXX
Identify Products

- Bankname: XX
- CCNumber: XXXXXXXX
- CCHolder: XX XXXX
- CCExpire: X / XXXX
- CVV2: XX
- Vorname: XX
- Nachname: YY
- Adress: XXXXX
- Stadt: XXXX
- PLZ: XXXX
- Land: XX
- Telefon: XXXXX-XXXXX
- E-mail: [email]victim@example.com[/email]
- Geburtsdatum: XX / XX / XXXX

Find repeated patterns

Check for verb

Tag parts of speech

Product
Profiling using writing style

Combine all messages of a member → Identify Product → Find members with enough texts

Extract features → Train classifier to model a member → Evaluate

JStylo
Language specific features perform better.

![Bar chart showing comparison between language specific and language independent features for different services.](chart.png)

- **L33tCrew (private)**
- **L33tCrew (public)**
- **Carders (private)**
- **Carders (public)**
- **Antichat (public)**
- **Blackhat (public)**

- **Language specific**
- **Language independent**
Accuracy is better in private messages.

Language specific

Language independent
Relaxed attribution

Exact attribution:

Who is this author?

Classifier

Alice ➔

This is Alice

This is Alice
Relaxed attribution

Who is this author?

Find top n possible authors

They could be Alice

Alice
Number of possible members vs. Accuracy for different communities:

- Antichat (public)
- Blackhat (public)
- Carders (private)
- Carders (public)
- L33tCrew (private)
- L33tCrew (public)

Accuracy climbs as the number of possible members increases, with different lines representing each community type.
Tracking members across forums

• Can a member’s posts on one forum identify him in other forums?
Tracking members across forums

• Approach:
  • Find members using email address
  • Train on one forum’s posts and test on another
Tracking members across forums

Common members

<table>
<thead>
<tr>
<th></th>
<th>Carders</th>
<th>L33tCrew</th>
</tr>
</thead>
<tbody>
<tr>
<td>Members with not enough text</td>
<td>39</td>
<td>83</td>
</tr>
<tr>
<td>Members with enough text</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Friday, December 28, 12
Cross forum result

Train on carders (39 members)

Train on L33tCrew (83 members)
Who are the possible suspects?

Find top 10 possible identities of Alice
Who are the possible suspects?

What do they tell us about Alice?

Find top 10 possible identities of Alice
Who are the possible suspects?

What do they tell us about Alice?

Find top 10 possible identities of Alice

Can we find alternate identities of Alice?
Who are the possible suspects?

If Bob is one of the suspects of Alice

Created a graph where each node= a user edge a to b if b is one of the suspects of a
Who are the possible suspects?

- Run this analysis of SpamIt forum associates chat conversation
- Because we had ground truth about duplicate accounts
Who are the possible suspects?

Bob

Alice

@alice

Alice2

ICQ:Alice
Discover Topic

- Topic analysis can identify predominant topic of the forum

- Why is it important?:
  - Automatically identify “interesting” subset of the data
  - Find relevant people
  - Understand trends
Discover Topic

- We used Latent Dirichlet Allocation (LDA) for identifying topic words.
How LDA works

I wonder what kind of products or services you could advertise using bulk mailing. I'm really new to bulk mailing, never tried it before. But I know that you'll get in trouble with your hosting company if you promote stuff that leads to your own server / hosting account. So how does this work???

Example post from Blackhat
How LDA works

I wonder what kind of products or services you could advertise using bulk mailing. I'm really new to bulk mailing, never tried it before. But I know that you'll get in trouble with your hosting company if you promote stuff that leads to your own server / hosting account. So how does this work???

Example post from Blackhat
Discover Topic

Find number of topics → Find topics and topics proportion → Find top 20 words per topic → Manual analysis
Discover Topic

Find number of topics
(200-300 on average)

Find topics and topics proportion

Find top 20 words per topic

Manual analysis
Topics discovered

Carders
- Anonymity services (web and phone): 19%
- Exploits: 14%
- Carding and other accounts: 11%
- Cardable: 14%
- Bankdrop: 10%
- Drugs: 5%
- Currency: PSC, UKASH, WMZ: 11%

L33tCrew
- Crypting services: 33%
- Anonymity services: 16%
- Carding: 26%
- Carding and other accounts: 12%
- Other accounts: 7%
- Exploits: 5%
- Anonymous phone: 11%
Topics discovered

Carders

Carding

Currency

- Anonymity services (web and phone)
- Exploits
- Carding and other accounts
- Cardable
- Bankdrop
- Drugs
- Currency: PSC, UKASH, WMZ

L33tCrew

Crypting services
- Anonymity services
- Carding
- Other accounts
- Exploits
- Anonymous phone
Topics discovered

Popular topics:
- Anonymity services (web and phone) 30%
- Exploits 11%
- Carding and other accounts 11%
- Cardable 19%
- Bankdrop 14%
- Drugs 10%
- Currency: PSC, UKASH, WMZ 5%
- L33tCrew 12%
- Crypting service 7%
- Other accounts 5%
- Carders 16%
- Anonymity service 26%
- Carding 33%

Friday, December 28, 12
Topics discovered

Antichat

- Web money: 54%
- Exploits: 33%
- Phone and SMS: 33%
- Captcha solving: 6%
- SEO blackhat tools: 9%
- Password cracking: 25%

Blackhat

- SEO blackhat: 49%
- SEO for blogs: 18%
- SEO for youtube: 16%
- Captcha solving: 7%
- Buy followers and friends: 11%
Next...

• Challenges
• Limitations
• Future work
• Conclusions
• Tools developed in our lab
CHALLENGES

• Microtext
• Multilingual text
• Different types of product information in text
• Users with multiple accounts
Challenges of microtext

- Short writings
  “1x Brazzers heut morgen gings noch”
Challenges of microtext

• Short writings
  “1x Brazzers heut morgen gings noch”

• Informal and conversational style
  “LOL............ nice post”
Challenges of multilingual text

• Require multilingual features in machine learning
  • Language-specific POS tagger
  • Function words
Challenges of multilingual text

• Many authorship tools are designed for English
• Translated text gives better results in identifying members
Challenges of multilingual text

• We translated Carders public to see how translation affects the accuracy
Challenges of multilingual text

• We translated Carders public to see how translation affects the accuracy
Challenges of multilingual text

- Highest accuracy achieved through the translated dataset that used English features
Challenges of translating multilingual text

- Large dataset requires automatic language detection for batch translations
- Low quality translations because of microtext properties
Challenges of multilingual text

- Automatic language detection is hard
Challenges of multilingual text

• Automatic language detection is hard
Challenges of multilingual text

- Automatic language detection is hard
Challenges of multilingual text

• Low quality translations
Challenges of multilingual text

- Low quality translations
Challenges of multilingual text

- Low quality translations
Challenges of multilingual text

• Low quality translations

Translation:

Cracker MSN: MSN is going to help you break the address you want to break a programdır. Oluşturduğunuz worldlistelerle password attempt can break through.
Challenges of multilingual text

- Low quality translations

Translation:

Cracker MSN: MSN is going to help you break the address you want to break a programdir. Olusturduğunuz worldlistelerle password attempt can break through.
Challenges of multilingual text

• Low quality translations

Translation:

Not English

Cracker MSN: MSN is going to help you break the address you want to break a programdır. Oluşturduğunuz worldlistelerle password attempt can break through.
Challenges of product information in text

• Problematic for language independent features such as character n-grams
• Adds noise to the model for each author
• Messages contain both conversation and product information

Message:
халявщики просыпайтесь..этот раздел не умрет ;) (наверно) :D кто забрал,отпишитесь)
1234;xxxx 1234@rambler.ru
1235;yyyy 1235@rambler.ru
1236;zzzz 1236@rambler.ru
Challenges of product information in text

• Problematic for language independent features such as character n-grams
• Adds noise to the model for each author
• Messages contain both conversation and product information

Message:
халявщики просыпайтесь..этот раздел не умрет ;) (наверно) :D кто забрал,отпишитесь)
1234;xxxx 1234@rambler.ru
1235;yyyy 1235@rambler.ru
1236;zzzz 1236@rambler.ru
Challenges of product information in text

• Problematic for language independent features such as character n-grams
• Adds noise to the model for each author
• Messages contain both conversation and product information

Message:
халявщики просыпайтесь..этот раздел не умрет ;) (наверно) :D кто забрал,отпишитесь)
1234;xxxx 1234@rambler.ru
1235;yyyy 1235@rambler.ru
1236;zzzz 1236@rambler.ru
Challenges of product information in text

• We have this huge dataset
Challenges of product information in text

• We have this huge dataset
• We need to detect products so that we can analyze the conversational text
Challenges of product information in text

• We have this huge dataset
• We need to detect products so that we can analyze the conversational text
• How can you build a method to detect different types of product information
Challenges of product information in text

• We have this huge dataset
• We need to detect products so that we can analyze the conversational text
• How can you build a method to detect different types of product information
• We consider a text pattern that doesn’t contain verbs as product information. This is done with a POS-tagger.
Types of products

- Exploits
- Copyright infringement
- Credit cards
- Bank accounts
- E-mail accounts
- Online accounts
- Bankdrops
- Shipping/delivery services
- Drugs
# Exploit coded by “Alice” and “Bob”

```
use LWP::UserAgent;
$ua = new LWP::UserAgent;
$ua->agent("Mosiac 1.0" . $ua->agent);
if (!$ARGV[0]) {$ARGV[0] = '';}
if (!$ARGV[3]) {$ARGV[3] = '';}
my $path = $ARGV[0] . '/index.php?
act/Login&CODE=autologin';
my $user = $ARGV[1]; # userid to jack
my $iver = $ARGV[2]; # version 1 or 2

....
if (!$ARGV[2])
{print "The type of the file system is NTFS.

WARNING, ALL DATA ON NON-REMOVABLE DISK";
print "WARNING, ALL DATA ON NON-REMOVABLE DISK
";}
...
```
Copyright infringement

Style: Format: MP3, 160 kbps Size: 50,7 Mb Country: USA

01. "Flat Line"
02. "6 6 Sick"
03. "Addiction" (featuring Zakk Wylde)
04. "No Regrets"
05. "My Funeral"
06. "We Are"
07. "Dirty World"
08. "Interlude"
09. "Violence"
10. "Best for Me"
11. "Bloodless"
12. "Scorn"
13. "Rebel Yell" (Billy Idol cover)
14. "I Don't Give a..."
15. "Die, Boom, Bang, Burn, F*ck"
16. "Nothing for Me Here"

Download:
http://rapidshare.com/files/123456789/NR-copyrightportal.ru.rar
http://depositfiles.com/files1a2b3c4d5e
Credit Card

DE CCV MasterCard
I have got here a Master Card Germany checked, but I can't need it anymore...

*CardholderName:Alice Smith
*FirstName:Alice
*LastName:Smith
*Address:Alice’s Street and number
*ZIP:99999
*City:Alice’s city
*Country:DE
*Phone:12345678900
*Email:alice_smith@xxx.net
*1234123412341234
*012
*0123
Bank Account

Data: Fri Dec 28, 2012 11:11 pm
Login: alice_smith@xxx.net
Parola: Alice’s_password
First Name: Alice
Last Name: Smith
Card Type: Visa
Bank Name: Citigroup Smith Barney
CC Number: 4321432143214321
Month: 01
Year: 2015
CVV2: 215
PIN: 2345
More accounts

====================================================================
Software          : Windows Live Messenger
Protocol          : MSN Messenger
User                : AliceSmith@ggg.de
Password        : Alice’s_password
====================================================================

====================================================================
Software           : ICQ Lite/2003
Protocol            : ICQ
User                  : 123454321
Password          : Alice’s_password
====================================================================

====================================================================
Name            : AliceSmith
Application    : Hotmail/MSN
Email             : AliceSmith@ggg.de
Server           :
Type              : HTTP
User              : AliceSmith
Password      : Alice’s_password
Profile           :
====================================================================
Online accounts

paypal .................

254/alice_smith1@xxx.net/Alice's_password
253/alice_smith2@xxx.net/Alice's_password
252/alice_smith3@xxx.net/Alice's_password
251/alice_smith4@xxx.net/Alice's_password
250/alice_smith5@xxx.net/Alice's_password
249/alice_smith6@xxx.net/Alice's_password
248/alice_smith7@xxx.net/Alice's_password
247/alice_smith8@xxx.net/Alice's_password
246/alice_smith9@xxx.net/Alice's_password
245/alice_smith10@xxx.net/Alice's_password
244/alice_smith11@xxx.net/Alice's_password
243/alice_smith12@xxx.net/Alice's_password
242/alice_smith13@xxx.net/Alice's_password
241/alice_smith14@xxx.net/Alice's_password
240/alice_smith15@xxx.net/Alice's_password
239/alice_smith16@xxx.net/Alice's_password
238/alice_smith17@xxx.net/Alice's_password
237/alice_smith18@xxx.net/Alice's_password
236/alice_smith19@xxx.net/Alice's_password
Bankdrops

Bankdrop-tutorial

Was wird benötigt?

- Socks/VPN etc.
- 1 Fake Acc (in 2min. erledigt)
- 1 Fake Foto (in 2min. erledigt)
- 1 Fake E-Mail (in 2min. erledigt)
- Briefkastendrop
Shipping/Delivery Services

Once we receive note of order your selected equipment will be shipped to your indicated address no longer than 2-3 working days you will receive an email with shipping track and trace.

To place order email below.
Email: alice_smith@xxx.net
ICQ support: 123*123*123 or 321*321*321

SKIMMER PACKAGES AND PRICE LIST
SKIMMERS SOLD SEPARATELY

Atm Model:
Diebold / Wincor / Ncr (skimmer only)
Type: USB
Price: $2000

Atm Model:
Diebold / Wincor / Ncr (skimmer only)
Type: Bluetooth
Price: $2500
Skimmer that looks like anti-skimming device
Drugs

Message:
Verified Vendor of **Weed**

Message:
Intresse an **Drug Store** ?
weed
pep
mdma
Challenges caused by users with multiple accounts

- Same user with different IP and e-mail address opens multiple accounts to avoid being banned
Challenges caused by users with multiple accounts

- Same user with different IP and e-mail address opens multiple accounts to avoid being banned
- Difficult to identify multiple account holders
Challenges caused by users with multiple accounts

• Same user with different IP and e-mail address opens multiple accounts to avoid being banned
• Difficult to identify multiple account holders
• In supervised learning, we consider them as different users
Challenges caused by users with multiple accounts

• Same user with different IP and e-mail address opens multiple accounts to avoid being banned
• Difficult to identify multiple account holders
• In supervised learning, we consider them as different users
• Authorship attribution classification accuracy and social connection graphs suffer due to this lack of ground truth
Limitations

• The required text length is 5000 words
Limitations

- The required text length is 5000 words
- Forums with less well known languages anticipated
Limitations

• The required text length is 5000 words
• Forums with less well known languages anticipated
• Separate conversational data from product information for better stylometric analysis
Limitations

• The required text length is 5000 words
• Forums with less well known languages anticipated
• Separate conversational data from product information for better stylometric analysis
• Our method worked well but needs improvements
Limitations

• The required text length is 5000 words
• Forums with less well known languages anticipated
• Separate conversational data from product information for better stylometric analysis
• Our method worked well but needs improvements
• False-positive product detection:
Limitations

• The required text length is 5000 words
• Forums with less well known languages anticipated
• Separate conversational data from product information for better stylometric analysis
• Our method worked well but needs improvements
• False-positive product detection:

Msn Piç : Sinir olduğunuz insanlara ister kendi zevkiniz ister programda bulunan küfürleri yollayabilirsiniz.Sizden fazla hızlı küfür edemez :)
Limitations

• The required text length is 5000 words
• Forums with less well known languages anticipated
• Separate conversational data from product information for better stylometric analysis
• Our method worked well but needs improvements
• False-positive product detection:

Msn Piç : Sinir olduğunuz insanlara ister kendi zevkiniz ister programda bulunan küfürleri yollayabilirsiniz. Sizden fazla hızlı küfür edemez :)
Conclusions

• We applied stylometric analysis to a huge real world dataset.
Conclusions

• We applied stylometric analysis to a huge real world dataset.
• This has been very rarely done
Conclusions

• We applied stylometric analysis to a huge real world dataset.
• This has been very rarely done
• Short leetspeak cannot be easily translated.
Conclusions

• We applied stylometric analysis to a huge real world dataset.
• This has been very rarely done
• Short leetspeak cannot be easily translated.
• Applying stylometry to big and semi-structured data is a difficult thing.
Conclusions

• We applied stylometric analysis to a huge real world dataset.
• This has been very rarely done
• Short leetspeak cannot be easily translated.
• Applying stylometry to big and semi-structured data is a difficult thing.
• We raised our accuracy to authorship attribution quality on a dataset that is not clean
Conclusions

• We applied stylometric analysis to a huge real world dataset.
• This has been very rarely done.
• Short leetspeak cannot be easily translated.
• Applying stylometry to big and semi-structured data is a difficult thing.
• We raised our accuracy to authorship attribution quality on a dataset that is not clean.
• Stylometry helps us identify suspects and predominant topics.
Conclusions

• We applied stylometric analysis to a huge real world dataset.
• This has been very rarely done
• Short leetspeak cannot be easily translated.
• Applying stylometry to big and semi-structured data is a difficult thing.
• We raised our accuracy to authorship attribution quality on a dataset that is not clean
• Stylometry helps us identify suspects and predominant topics
• We minimize manual analysis time
Future work

• Use more user-specific features and temporal information
Future work

- Use more user-specific features and temporal information
- Add topic information with authorship information
Future work

• Use more user-specific features and temporal information
• Add topic information with authorship information
• Identify multiple account holders
Future work

• Use more user-specific features and temporal information
• Add topic information with authorship information
• Identify multiple account holders
• Combine interaction from different media (IRC chat logs with forums)
Future work

- Use more user-specific features and temporal information
- Add topic information with authorship information
- Identify multiple account holders
- Combine interaction from different media (IRC chat logs with forums)
- Completely automate the process of identifying users with sufficient text from the datasets and perform topic and authorship analysis
Summary

• Profiling members of the underground economy
Summary

• Profiling members of the underground economy
• Product identification
Summary

• Profiling members of the underground economy
  • Product identification
• Discovering topics being discussed by these members
Summary

• Profiling members of the underground economy
  • Product identification
• Discovering topics being discussed by these members
• Interaction network analysis
Summary

• Profiling members of the underground economy
  • Product identification
• Discovering topics being discussed by these members
• Interaction network analysis
• Challenges
Summary

• Profiling members of the underground economy
  • Product identification
• Discovering topics being discussed by these members
• Interaction network analysis
• Challenges
• Limitations
Summary

• Profiling members of the underground economy
  • Product identification
• Discovering topics being discussed by these members
• Interaction network analysis
• Challenges
• Limitations
• Conclusions
Summary

• Profiling members of the underground economy
  • Product identification
• Discovering topics being discussed by these members
• Interaction network analysis
• Challenges
• Limitations
• Conclusions
• Future Work
JStylo

Our authorship attribution framework, powered by JGAAP and WEKA

Released in 28C3:

http://events.ccc.de/congress/2011/Fahrplan/events/4781.en.html
Anonymouth

Our authorship anonymization framework, powered by JStylo

Released in 28C3:

http://events.ccc.de/congress/2011/Fahrplan/events/4781.en.html
Anonymouth

- Your writing style can give you away
Anonymouth

• Your writing style can give you away
• Anonymouth identifies changes required for document anonymization relative to a corpus
Anonymouth

• Your writing style can give you away
• Anonymouth identifies changes required for document anonymization relative to a corpus
• Assists the user making necessary changes accordingly
https://github.com/psal/JStylo-Anonymouth

JStylo + Anonymouth = JSAN
OPEN SOURCE IN GIT

https://psal.cs.drexel.edu/index.php/JStylo-Anonymouth
Questions?

• Sadia Afroz
  • sa499@cs.drexel.edu
• Aylin Caliskan Islam
  • ac993@cs.drexel.edu
• Ariel Stolerman
  • ams573@cs.drexel.edu
• Damon McCoy
  • mccoy@cs.gmu.edu
• Rachel Greenstadt
  • greenie@cs.drexel.edu
• Research at PSAL Drexel
  • https://psal.cs.drexel.edu/