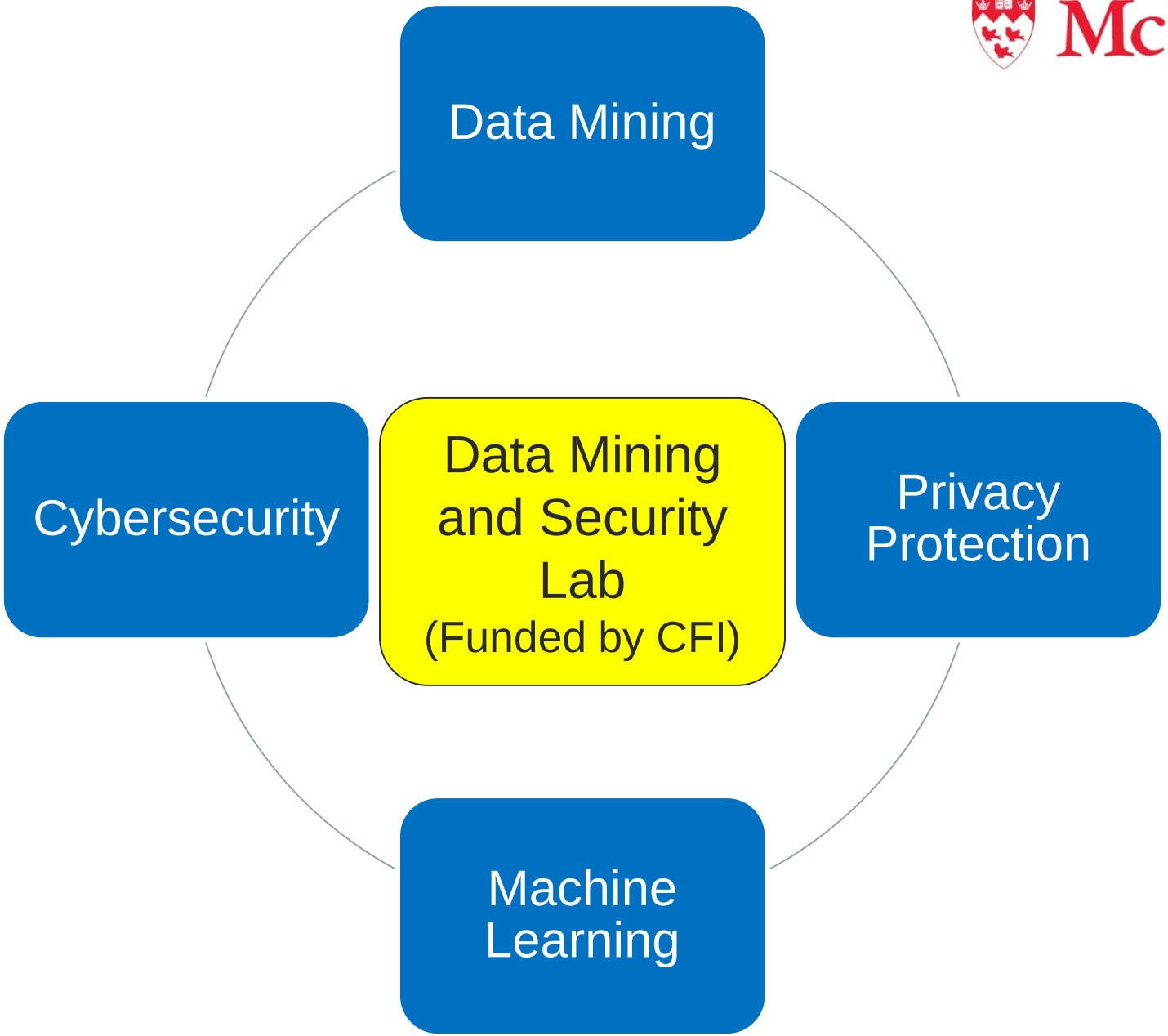
Data Mining and Machine Learning for Authorship and Malware Analyses

Benjamin Fung
Data Mining and Security Lab
School of Information Studies
McGill University
Canada

June 21st, 2019 @ Carleton University

McGill


Data Mining
Cybersecurity
Data Mining and Security Lab (Funded by CFI)
Privacy Protection
Machine Learning

McGill

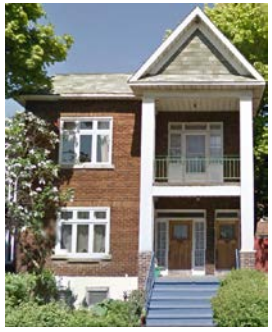

Healthcare
Law Enforcement
National Defense
Energy
Transportation

Agenda

- Authorship analysis for crime investigation
- Assembly code mining for malware analysis
- New ML project on neuroscience

Authorship Analysis for Crime Investigation

Real-life Criminal Case (Scam)



Homestay se

armela in US

Asked for C$2800. Received C$4500.

Anthony in Canada

SCOTT'S REAL ESTATE LIMITED PARTNERSHIP
181 BAY STREET, SUITE 2300
CT TOWER, BCE PLACE
TORONTO, ONTARIO M5J 2S1

BANK OF MONTREAL
ONE FIRST CANADIAN PLACE
TORONTO, ONTARIO M5X 1H3

1783

1783

Date 0 5 1 5 _ _ _ _
M M D D Y Y Y Y

Four Thousand Five Hundred and Zero Cents

$ *****4,500.00

PAY

TO THE ORDER OF BENJAMIN FUNG

SCOTT'S REAL ESTATE LIMITED PARTNERSHIP

PER

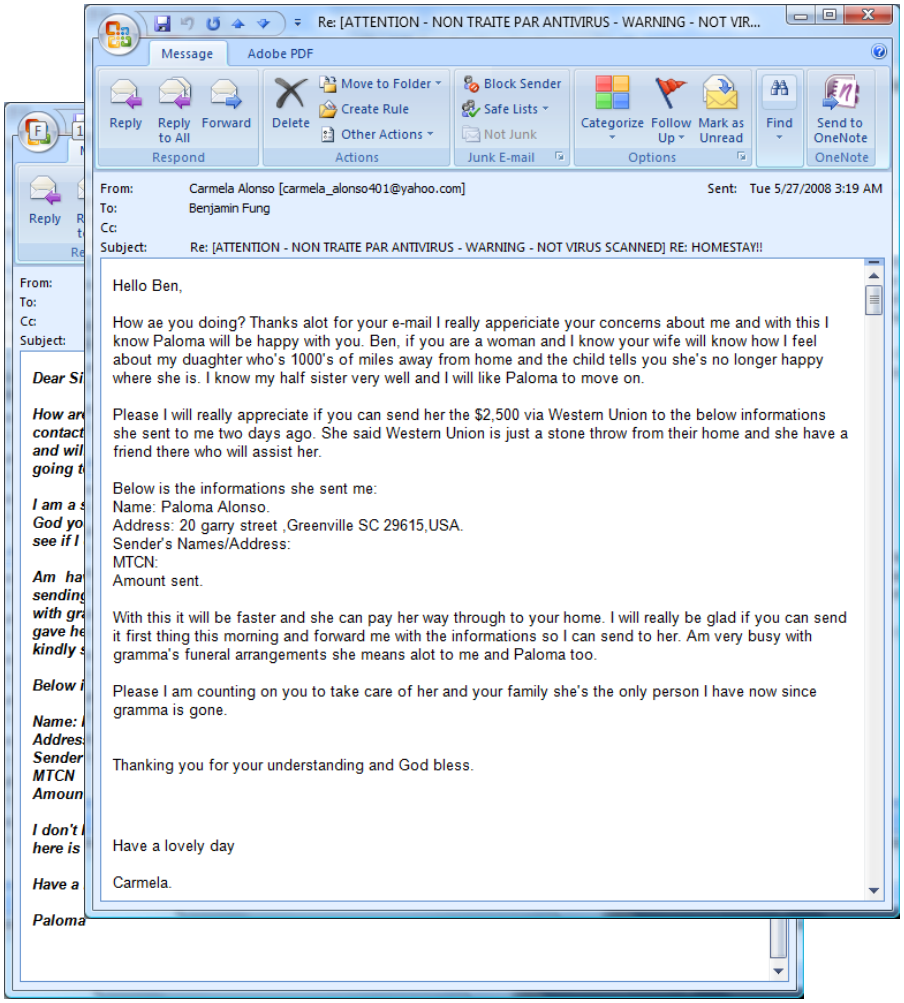
PER

⑈001783⑈ ⑆00022⑈001⑆ ⑆501⑈883⑈

Motivating Example: Cybercrime

- Cell phone number of Anthony:
- 15 e-mails from “Carmela”
- A counterfeit cheque

Anthony



Authorship Identification: Who is behind the keyboard?

Digital writings: emails, blogs, tweets, forums, etc.

```
Return-Path: <melody@covingtoninnovations.com>
Received: from spgw1.servdns.com [65.163.13.5] by smail4.servdns.com with SMTP;
  Sun, 13 Jan 2008 19:59:57 -0500
Received: from fmailhost02.isp.att.net (fmailhost02.isp.att.net [204.127.217.102])
      by spgw1.servdns.com (Sectorlink) with ESMTP id AA8DB300097
      for <mc@covingtoninnovations.com>; Sun, 13 Jan 2008 19:58:13 -0500 (EST)
Received: from hokusai (adsl-224-168-165.asm.bellsouth.net[74.224.168.165])
       by isp.att.net (frfwmhc02) with SMTP
       id <20080114005830H0200afj55e>; Mon, 14 Jan 2008 00:58:30 +0000
X-Originating-IP: [74.224.168.165]
From: "Melody Covington" <melody@covingtoninnovations.com>
To: <melody@maxcharge.com>,
      '"Michael A. Covington"' <mc@covingtoninnovations.com>
Subject: Appointments for the coming week
Date: Sun, 13 Jan 2008 19:58:29 -0500
Organization: Covington Innovations
Message-ID: <001101c85648$94774e60$6801a8c0@Hokusai>
MIME-Version: 1.0
Content-Type: multipart/alternative;
      boundary="----=_NextPart_000_0012_01C8561E.ABA14660"
X-Mailer: Microsoft Office Outlook 11
X-MimeOLE: Produced By Microsoft MimeOLE V6.00.2900.3198
Thread-Index: AchWSJPQySP0K1HFSpSwLo/S9GWHQA==
X-servdns-MailScanner-Information: Please contact the ISP for more information
X-servdns-MailScanner: Found to be clean
X-servdns-MailScanner-From: melody@covingtoninnovations.com
```

"RECEIVED" LINES show how message entered the Internet. Last one or two are most informative. Some may be fake.

"FROM" LINE is address given by the sender; may be totally false.

LINES THAT START WITH X are comments added by software; may be true or false.

Authorship Identification

Candidate X

Candidate Y

Candidate Z

?

?

?

Actual author of given anonymous text

Handcrafted Stylometric Features

From Fingerprint to Writeprint

Stylometric Features

- Lexical features
- Syntactic features
- Structural features
- Content-specific features
- Idiosyncratic features

- Word length
- Vocab. Richness
- Digit/cap distribution

Handcrafted Stylometric Features

From Fingerprint to Writeprint

Stylometric Features

- Lexical features
- Syntactic features
- Structural features
- Content-specific features
- Idiosyncratic features

- Punctuations
- Function words "of", "to", "as"…

Handcrafted Stylometric Features

From Fingerprint to Writeprint

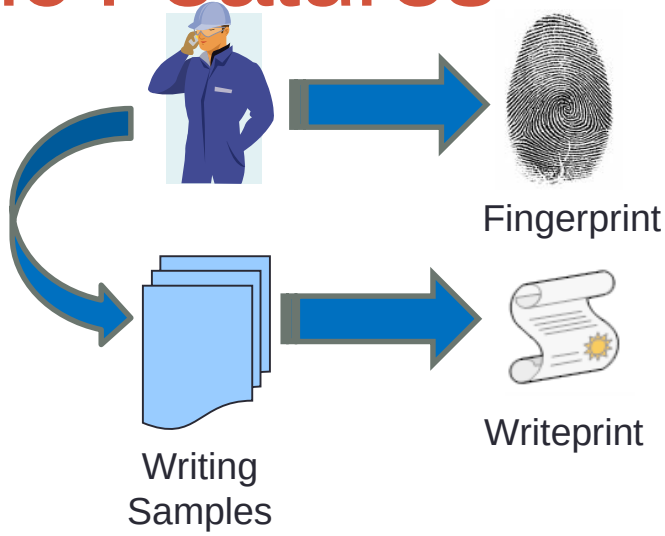
Stylometric Features

- Lexical features
- Syntactic features
- Structural features
- Content-specific features
- Idiosyncratic features

- Sentence length
- Paragraph length
- Separators b/w paragraphs

Handcrafted Stylometric Features

From Fingerprint to Writeprint



Stylometric Features

- Lexical features
- Syntactic features
- Structural features
- Content-specific features
- Idiosyncratic features

- Domain specific keywords
- Special characters

Handcrafted Stylometric Features

From Fingerprint to Writeprint

Stylometric Features

- Lexical features
- Syntactic features
- Structural features
- Content-specific features
- Idiosyncratic features

- Spelling and grammatical mistakes

Traditional Classifiers

1 Naïve Bayesian

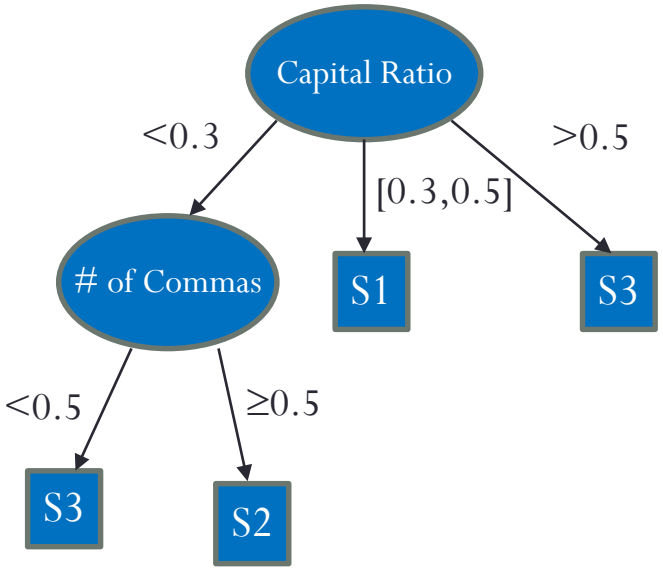
$$P(c_i \mid \vec{d}_j) = \frac{P(c_i)P(\vec{d}_j \mid c_i)}{P(\vec{d}_j)}$$

Efficient

Pitfalls:
Low classification accuracy in authorship analysis

(Sebastiani 2002, Diederich et al. 2003, Dong et al. 2006)

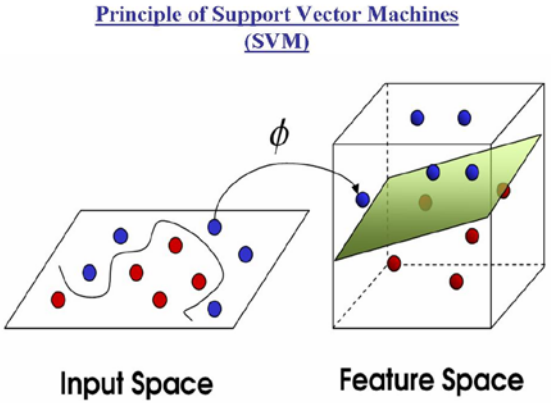
2 Decision Trees



Efficient, interpretable results

Pitfalls: Accuracy is comparable to SVM.
(Zhao & Zobel 2005, Sabastiani 2002)

3 Support Vector Machines

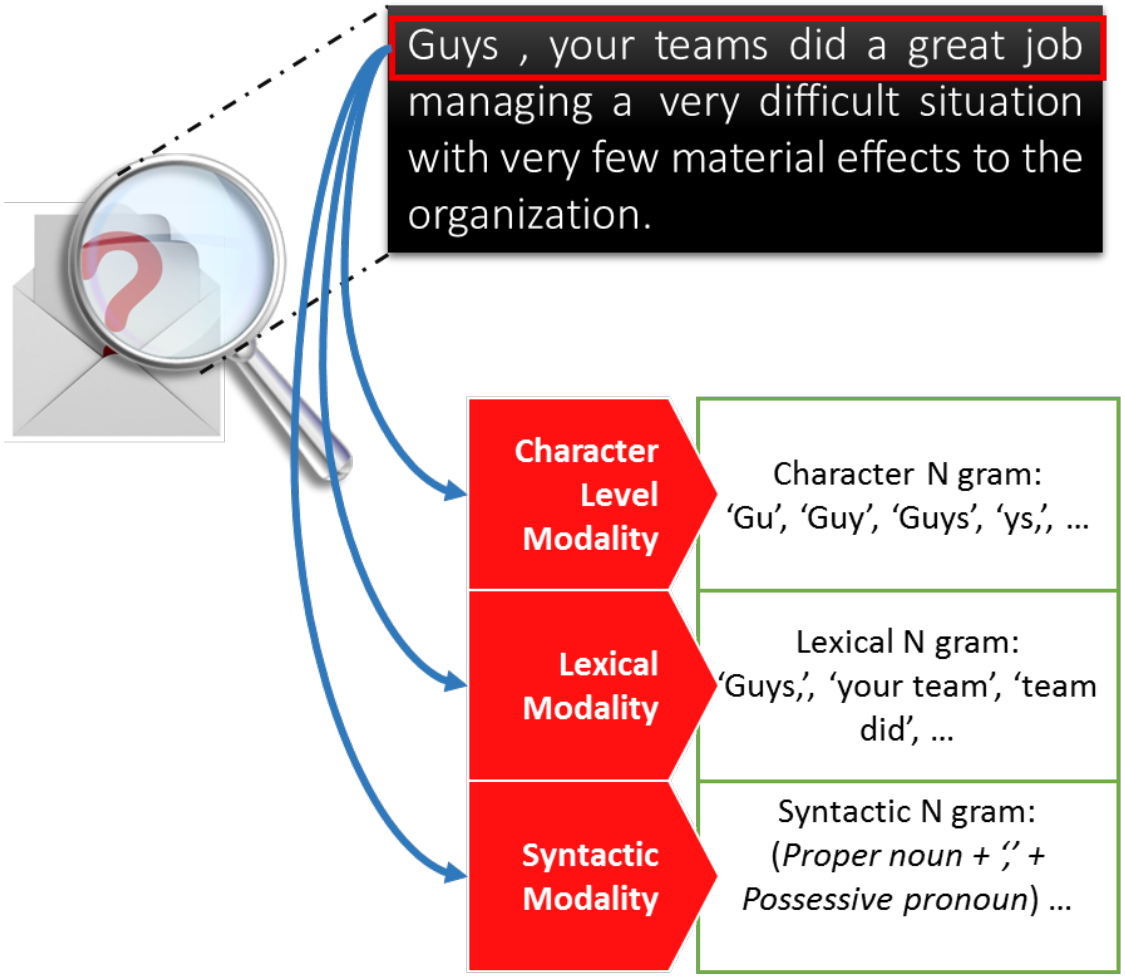


Reasonably accurate

Pitfalls:
Black box (difficult to interpret)

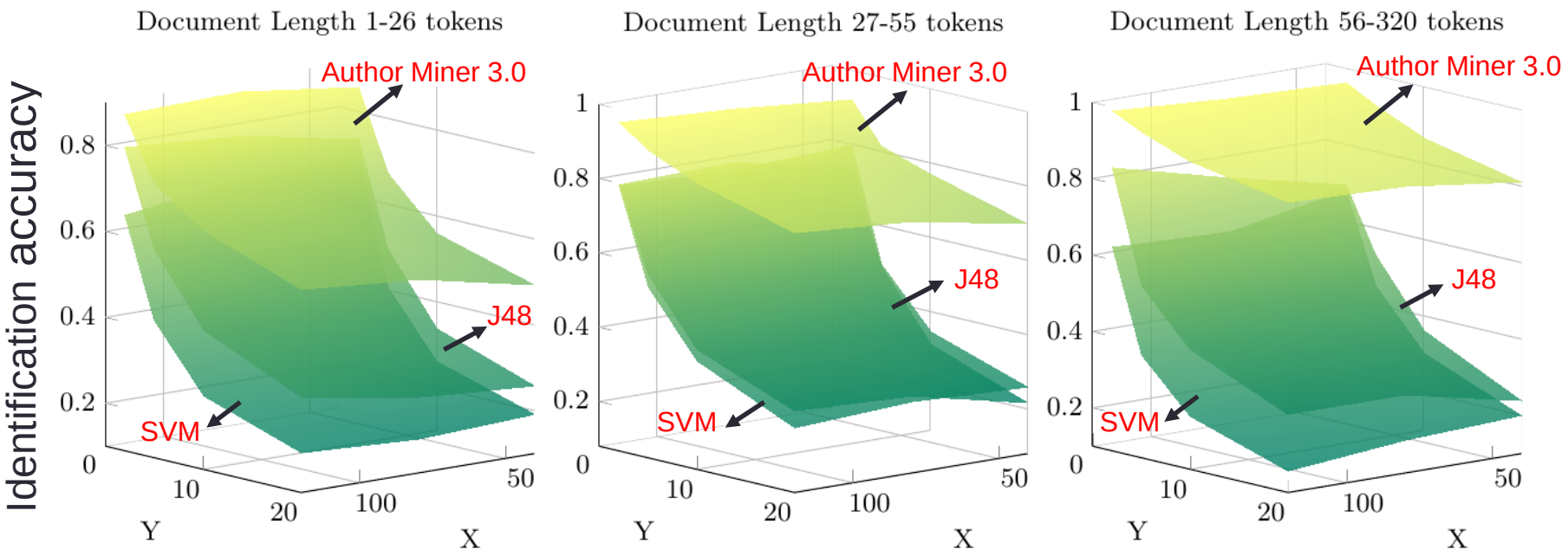
(de Vel et al. 2000)

AuthorMiner v3.0



- Calculate Affinity
- Estimate Confidence
- Visualize discrimination of affinity for each linguistic evidence unit.

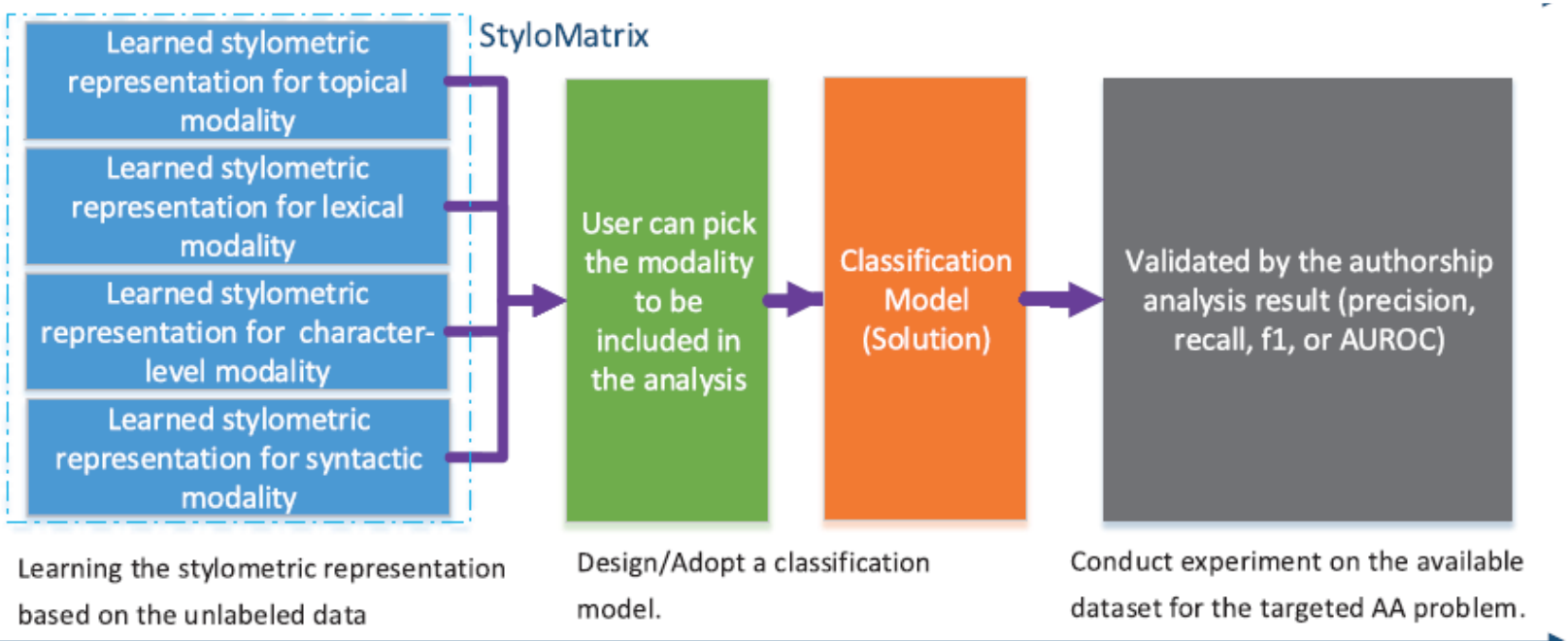
Performance
AuthorMiner 3.0



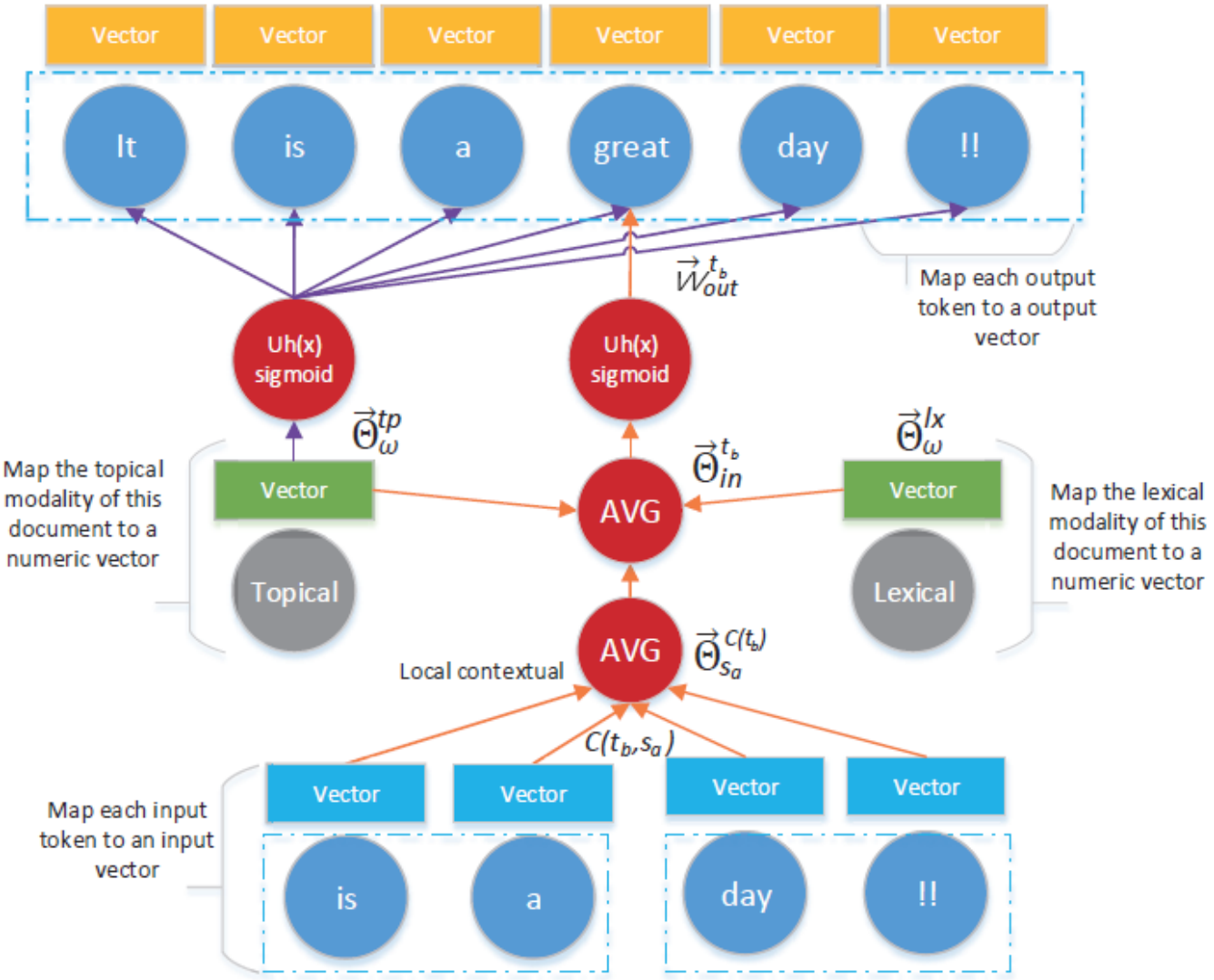
- X: number of writing samples per candidate (class)
- Y: number of candidates (classes)
- Dataset: Enron email dataset

Published in ACM TISSEC 2015

StyloMatrix: Learning Stylometric Representation



topic-lexic-2-vec



Published in IEEE Transactions on Cybernetics (2019)

Authorship Identification

- IMDb62 dataset: 62 profile users, each wrote 1,000 reviews
- Representation learning + logistic regression

Vector size	d=200	d=300	d=400	d=500	d=600
Topical+Lexical	0.9032	0.9209	0.9310	0.9379	0.9436
Topical	0.8327	0.8665	0.8779	0.8927	0.9028
Lexical	0.7369	0.7793	0.7979	0.8005	0.8037
Character	0.7185	0.7283	0.7330	0.7348	0.7298
Syntactic	0.3894	0.5104	0.5352	0.5828	0.6009

Cross-validation accuracy

Authorship Identification

- IMDb62 dataset: 62 profile users, each wrote 1,000 reviews

Model	Accuracy
[Lexical+Topical]*	**0.972**
Typed-n-gram [6]	0.937
[Topical]*	0.930
Token SVM [14]†	0.925
DADT-P [14]†	0.918
w2v-skigram+cbow	0.916
LSA	0.909
PV-DBOW+PV-DM	0.900
AT-P [14]†	0.896
Static+1000-n-gram	0.870
LDA+Hellinger-S [35]†	[0.80, 0.85]
Imposters (KOP)†	[0.70, 0.75]
[Lexical]*	0.742
[Character]*	0.733
LDA+Hellinger-M [35]†	< 0.70
LDA	0.677
[Syntactic]*	0.601

Authorship Verification

THE PAN2014 AUTHORSHIP VERIFICATION DATASET. THE NUMBER IN ROUND BRACKETS IS THE STANDARD DEVIATION.

Training	#Problems	#Docs	#Tokens	Tokens per doc
Dutch-Essays	96	192	123,713	644 (551)
Dutch-Reviews	100	200	25,416	127 (66)
English-Essays	200	400	694,477	1,736 (1372)
English-Novels	100	200	723,412	3,617 (3973)
Greek-Articles	100	200	616,497	3,082 (2283)
Spanish-Articles	100	200	767,916	3,839 (2639)
Testing	**#Problems**	**#Docs**	**#Tokens**	**Tokens per doc**
Dutch-Essays	96	192	128,179	667.59 (522)
Dutch-Reviews	100	200	26,169	130.85 (81)
English-Essays	200	400	671,056	1,677 (1352)
English-Novels	200	400	2,831,531	7,078 (5091)
Greek-Articles	100	200	646,361	3,231 (2395)
Spanish-Articles	100	200	755,929	3,779 (2622)

AUROC

Approach	Dutch Essay	Dutch Review	English Essay	English Novel	Greek Article	Spanish Article	Avg.
Modality [Lexical+Topical]*	**0.998**	**0.744**	**0.887**	0.767	0.924	0.934	**0.881**
Modality [Lexical]*	**0.998**	0.658	0.885	**0.799**	**0.949**	**0.937**	0.871
PV-DBOW+PV-DM	0.979	0.670	0.847	0.738	0.934	0.859	0.838
Modality [Character]*	0.960	0.642	0.854	0.758	0.889	0.911	0.836
META-CLASSIFIER-PAN2014†	0.957	0.737	0.781	0.732	0.836	0.898	0.824
PV-DBOW	0.985	0.656	0.848	0.711	0.868	0.870	0.823

Authorship Characterization

Age

Gender

Political orientation

?

?

?

Characteristics of the author

Software Demonstration: StyloMatrix

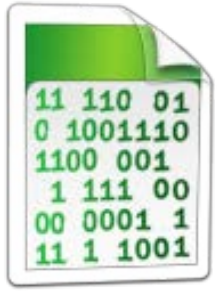
Assembly Code Mining for Reverse Engineering (Malware Analysis)

Reverse Engineering

```
void copy_block(deflate_state * s,
charf * buf, unsigned
          len, int header){

    bi_windup(s);
    s->last_eob_len = 8;
    if (header) {
        put_short(s, (ush)len);
        put_short(s, (ush)~len);

    }
}
```

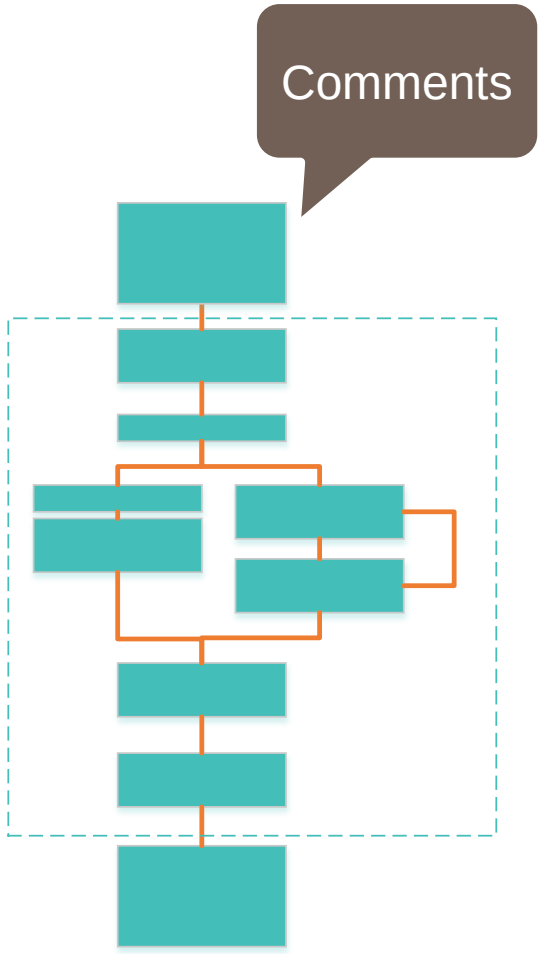



```
0xEB68 PUSH {R7,LR}
0xEB6A SUB SP, SP, #0x10
0xEB6C ADD R7, SP, #0
0xEB6E STR R0,
[R7,#0x10+s]
0xEB70 STR R1,
[R7,#0x10+buf]
0xEB72 STR R2,
[R7,#0x10+len]
0xEB74 STR R3,
[R7,#0x10+header]
0xEB76 LDR R0,
[R7,#0x10+s]
0xEB78 BL bi_windup
0xEB7C LDR R3,
[R7,#0x10+s]
0xEB7E ADD.W R3, R3,
#0x16A0
0xEB82 ADDS R3, #0x14
0xEB84 MOVS R2, #8
0xEB86 STR R2, [R3]
0xEB88 LDR R3,
[R7,#0x10+header]
0xEB8A CMP R3, #0
0xEB8C BEQ loc_EBFC
0xEB8E LDR R3,
[R7,#0x10+s]
0xEB90 LDR R2, [R3,#8]
0xEB92 LDR R3,
[R7,#0x10+s]
0xEB94 LDR R3, [R3,#0x14]
0xEB96 ADDS R0, R3, #1
0xEB98 LDR R1,
0xEBB8 LSRS R2, R2, #8
0xEBBA UXTH R2, R2
0xEBBC UXTB R2, R2
0xEBBE STRB R2, [R3]
0xEBC0 LDR R3,
[R7,#0x10+s]
0xEBC2 LDR R2, [R3,#8]
0xEBC4 LDR R3,
[R7,#0x10+s]
0xEBC6 LDR R3, [R3,#0x14]
0xEBC8 ADDS R0, R3, #1
0xEBCA LDR R1,
[R7,#0x10+s]
0xEBCC STR R0, [R1,#0x14]
0xEBCE ADD R3, R2
0xEBD0 LDR R2,
[R7,#0x10+len]
0xEBD2 UXTB R2, R2
0xEBD4 MVNS R2, R2
0xEBD6 UXTB R2, R2
0xEBD8 STRB R2, [R3]
0xEBDA LDR R3,
[R7,#0x10+s]
0xEBDC LDR R2, [R3,#8]
0xEBDE LDR R3,
[R7,#0x10+s]
0xEBE0 LDR R3, [R3,#0x14]
0xEBE2 ADDS R0, R3, #1
0xEBE4 LDR R1,
[R7,#0x10+s]
0xEBE6 STR R0, [R1,#0x14]
0xEBE8 ADD R3, R2
0xEBEA LDR R2,
0xEC0E LDR R3,
[R7,#0x10+buf]
0xEC10 ADDS R1
0xEC12 STR R1,
[R7,#0x10+buf]
0xEC14 LDRB R3,
0xEC16 STRB R3,
0xEC18 LDR R3,
[R7,#0x10+len]
0xEC1A SUBS R2
0xEC1C STR R2,
[R7,#0x10+len]
0xEC1E CMP R3,
0xEC20 BNE loc_
0xEC22 ADDS R7
0xEC24 MOV SP,
0xEC26 POP {R7,
```

Reverse Engineering

Reverse Engineering

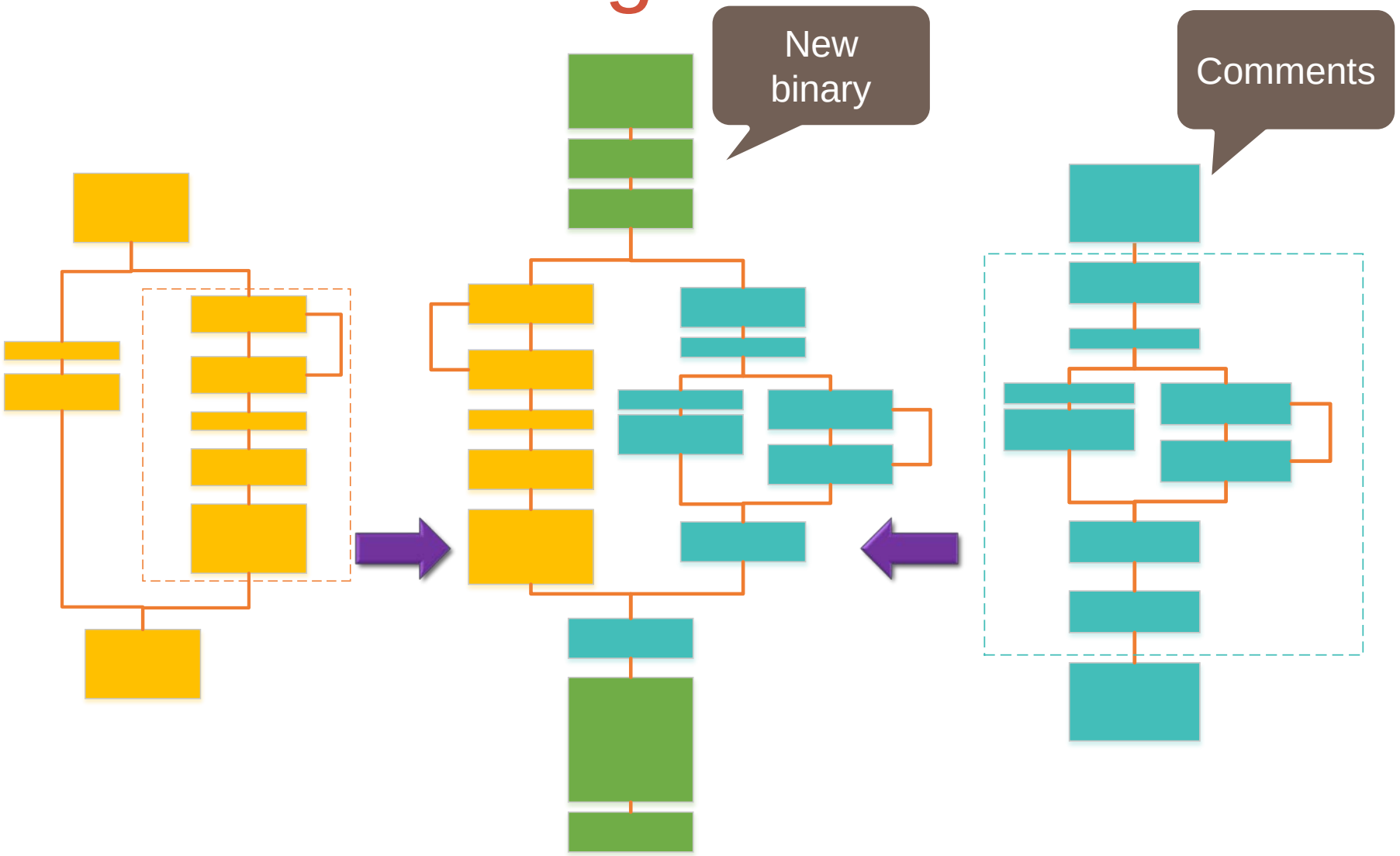
```
push    ebp
mov     ebp, esp
mov     eax, [ebp+arg_4]
push    ebx
push    esi
test    eax, eax
jz      loc_10002C2A
mov     esi, [ebp+arg_0]
test    esi, esi
jz      loc_10002C2A
mov     ebx, [eax+1Ch]
test    ebx, ebx
jz      loc_10002C2A
  ...
```



Assembly Code Function

Control Flow Graph

Kam1n0: Subgraph Clones



The original one

0x1FE69C0	PUSH ebp
0x1FE69C1	MOV ebp, esp
0x1FE69C3	MOV ecx, [ebp+arg_0]
0x1FE69C6	PUSH ebx
0x1FE69C7	MOV ebx, [ebp+arg_8]
0x1FE69CA	PUSH esi
0x1FE69CB	MOV esi, ecx
0x1FE69CD	AND ecx, 0FFFFh
0x1FE69D3	SHR esi, 10h
0x1FE69D6	CMP ebx, 1
0x1FE69D9	+JNZ loc_1FE6A0C

Type I: Exact clone

0x1FE69C0	PUSH ebp
0x1FE69C1	MOV ebp, esp
0x1FE69C3	MOV ecx, [ebp+arg_0]
0x1FE69C6	PUSH ebx
0x1FE69C7	MOV ebx, [ebp+arg_8]
0x1FE69CA	PUSH esi
0x1FE69CB	MOV esi, ecx
0x1FE69CD	AND ecx, 0FFFFh
0x1FE69D3	SHR esi, 10h
0x1FE69D6	CMP ebx, 1
0x1FE69D9	+JNZ loc_1FE6A0C

0x1FE69C0	PUSH ebp
0x1FE69C1	MOV ebp, esp
0x1FE69C3	MOV ecx, [ebp+arg_0]
0x1FE69C6	PUSH ebx
0x1FE69C7	MOV ebx, [ebp+arg_8]
0x1FE69CA	PUSH esi
0x1FE69CB	MOV esi, ecx
0x1FE69CD	AND ecx, 0FFFFh
0x1FE69D3	SHR esi, 10h
0x1FE69D6	CMP ebx, 1
0x1FE69D9	+JNZ loc_1FE6A0C

Type II: Syntactically equivalent

0x1FE05B0	PUSH ebp
0x1FE05B1	MOV ebp, esp
0x1FE05B3	MOV ecx, [ebp+arg_0]
0x1FE05B6	PUSH ebx
0x1FE05B7	MOV ebx, [ebp+arg_8]
0x1FE05BA	PUSH esi
0x1FE05BB	MOV esi, ecx
0x1FE05BD	AND ecx, 0FFFFh
0x1FE05B3	SHR esi, 10h
0x1FE05B6	CMP ebx, 1
0x1FE05B9	+JNZ loc_1FE05BC

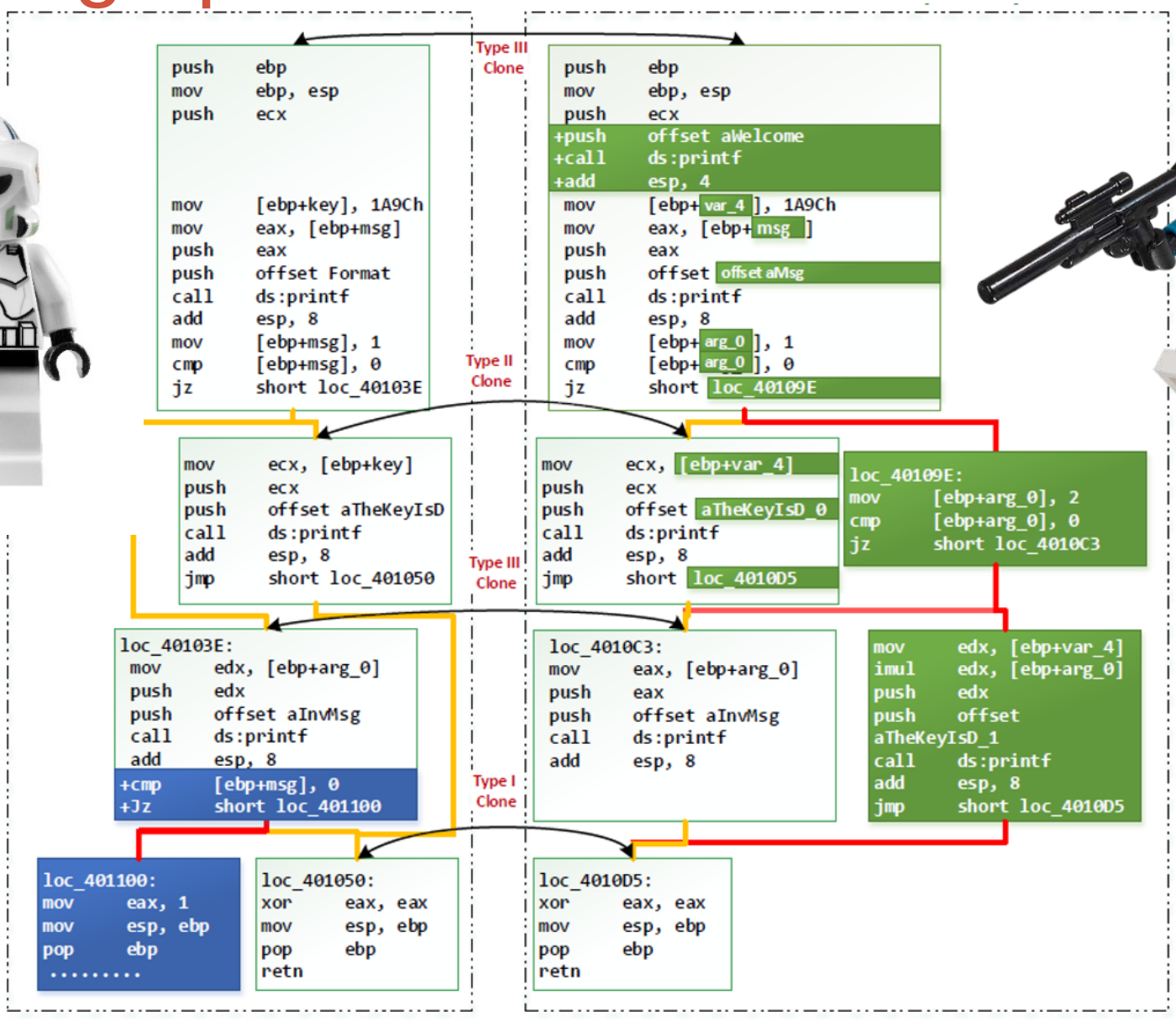
0x1FE69C0+	PUSH ebp
0x1FE69C1+	MOV ebp, esp
0x1FE69C3+	MOV eax, [ebp+msg_0]
0x1FE69C6+	PUSH edx
0x1FE69C7+	MOV edx, [ebp+msg_1]
0x1FE69CA+	PUSH esi
0x1FE69CB+	MOV esi, eax
0x1FE69CD+	AND eax, 0FFFFh
0x1FE69D3+	SHR esi, 10h
0x1FE69D6+	CMP edx, 1
0x1FE69D9+	+JNZ loc_1FE6A0C

Type III: Minor modification

0x1FE05B0	PUSH ebp
0x1FE05B1	MOV ebp, esp
+	
+	
0x1FE05B7	MOV ebx, [ebp+arg_8]
0x1FE05BA	PUSH esi
0x1FE05BB	MOV esi, ecx
0x1FE05BD	AND ecx, 0FFFFh
0x1FE05B3+	MOV eax, ecx
0x1FE05B6	SHR esi, 10h
0x1FE05B9	CMP ebx, 1
0x1FE05C1	+JNZ loc_1FE05BC

0x1FE69C0+	PUSH ebp
0x1FE69C1+	MOV ebp, esp
0x1FE69C3+	MOV eax, [ebp+msg_0]
0x1FE69C6+	PUSH edx
0x1FE69C7+	MOV edx, [ebp+msg_1]
0x1FE69CA+	PUSH esi
0x1FE69CB+	MOV esi, eax
0x1FE69CD+	AND eax, 0FFFFh
0x1FE69D3+	SHR esi, 10h
0x1FE69D6+	CMP edx, 1
0x1FE69D9+	+JNZ loc_1FE6A0C

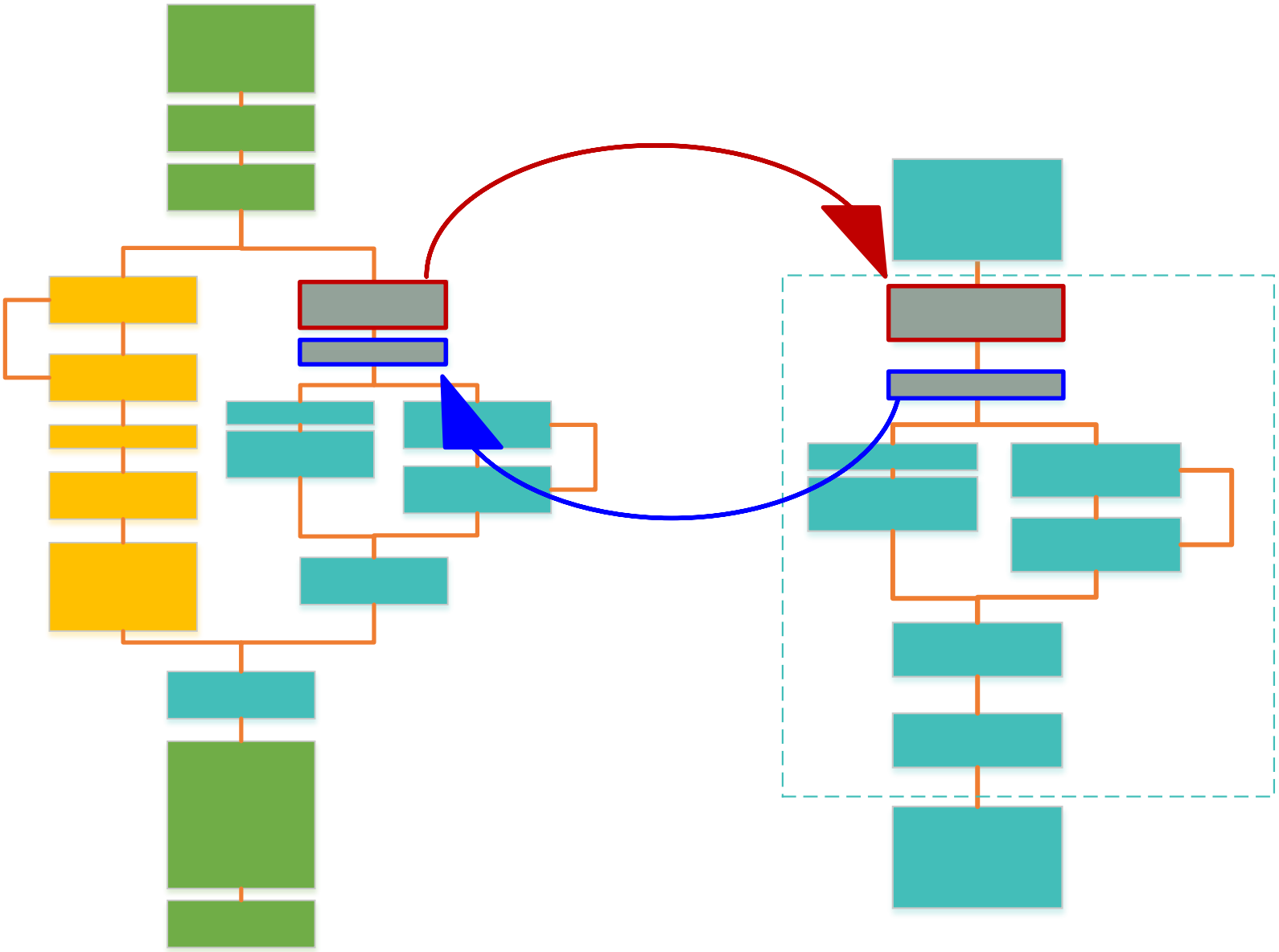
Subgraph clone



Challenges

- Accuracy: balance between recall and precision
- Efficiency and scalability: response time with terabytes of assembly code in repository
- Incremental update of the repository
- Interpretability and usability
- Flexibility: different CPU architectures, compilers, and optimization options

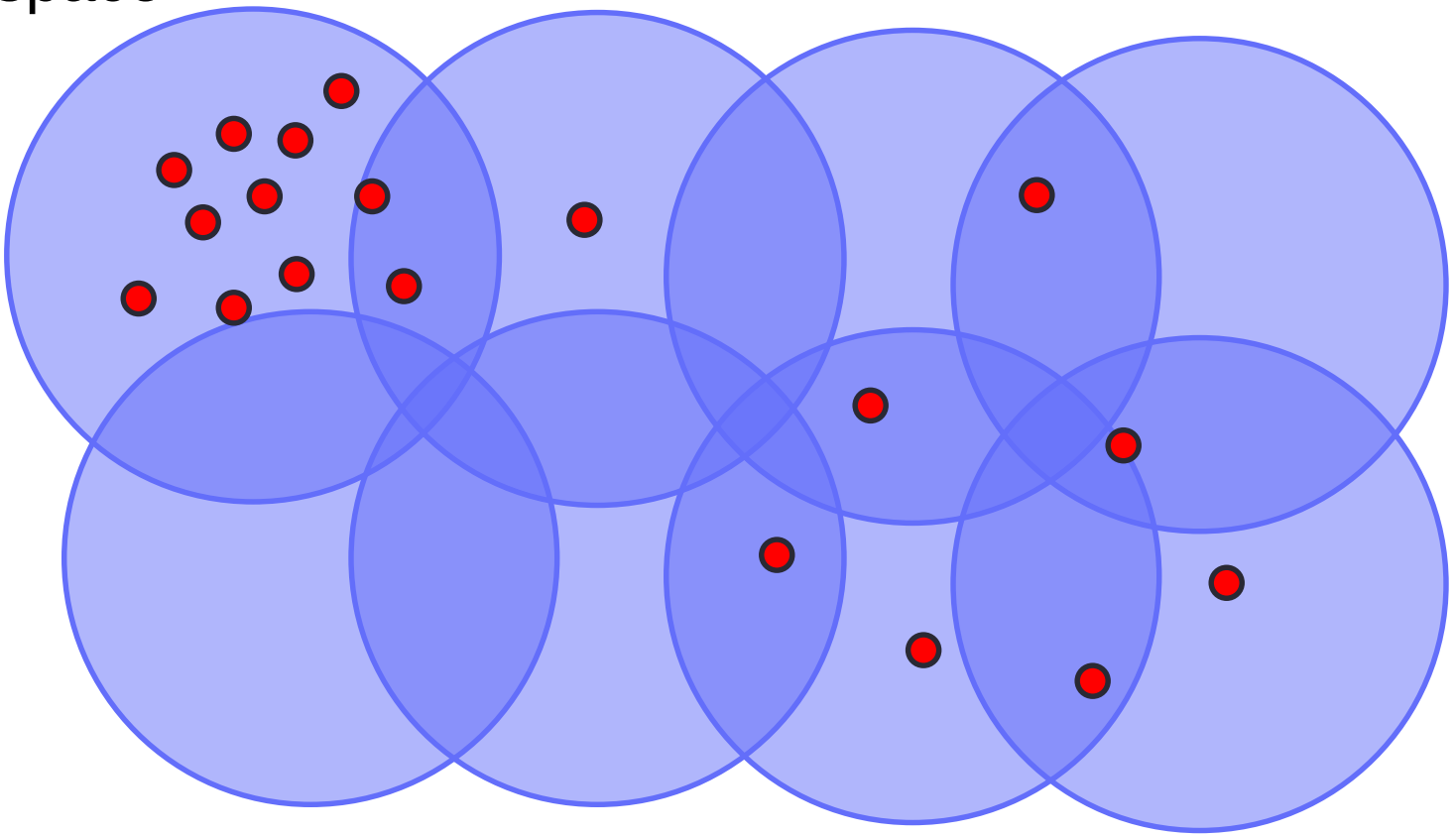
Finding Clones



Efficiency (*k*-NN problem)

Adaptive Locality Sensitive Hashing

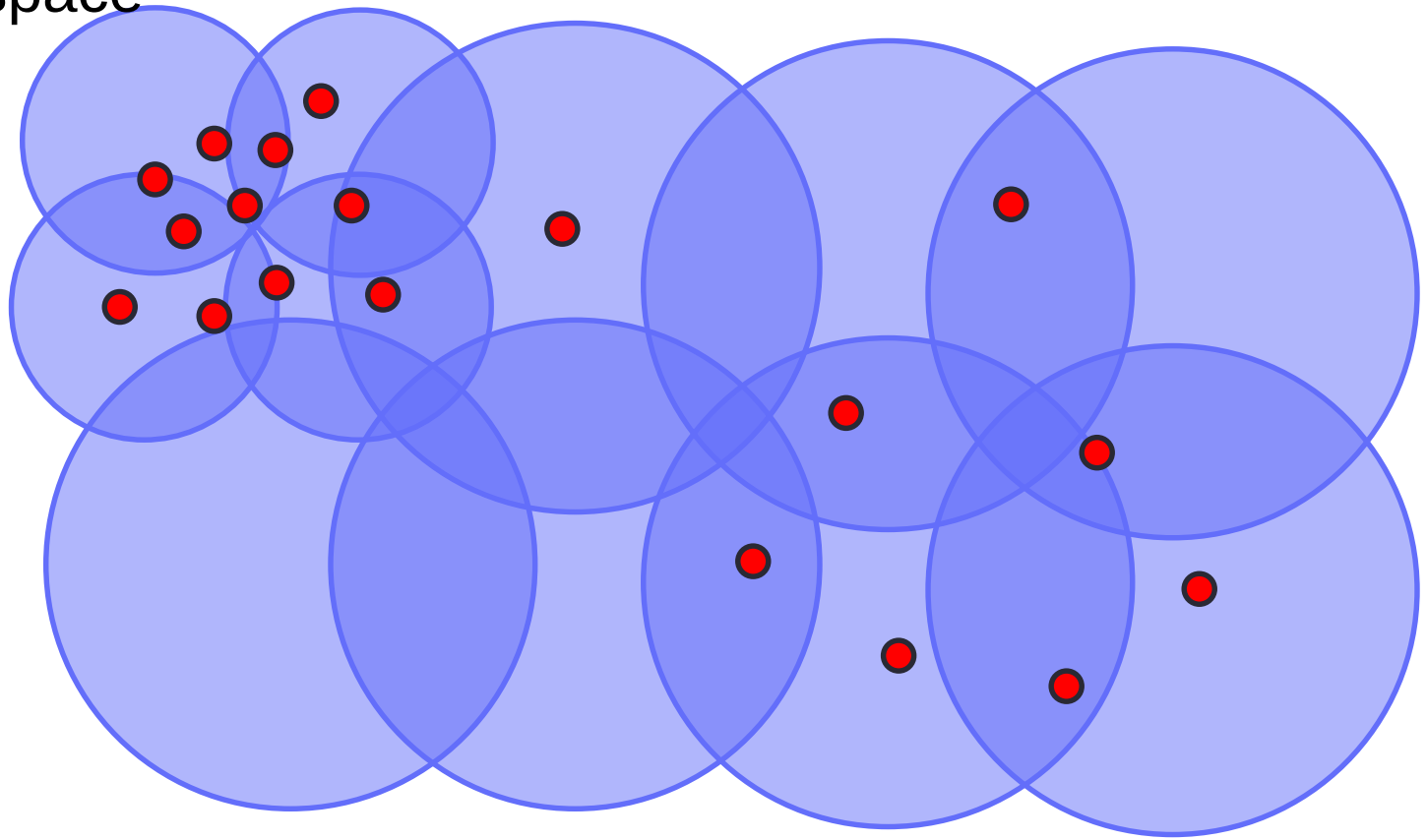
Unequal partitioning and dynamic hashing for cosine space



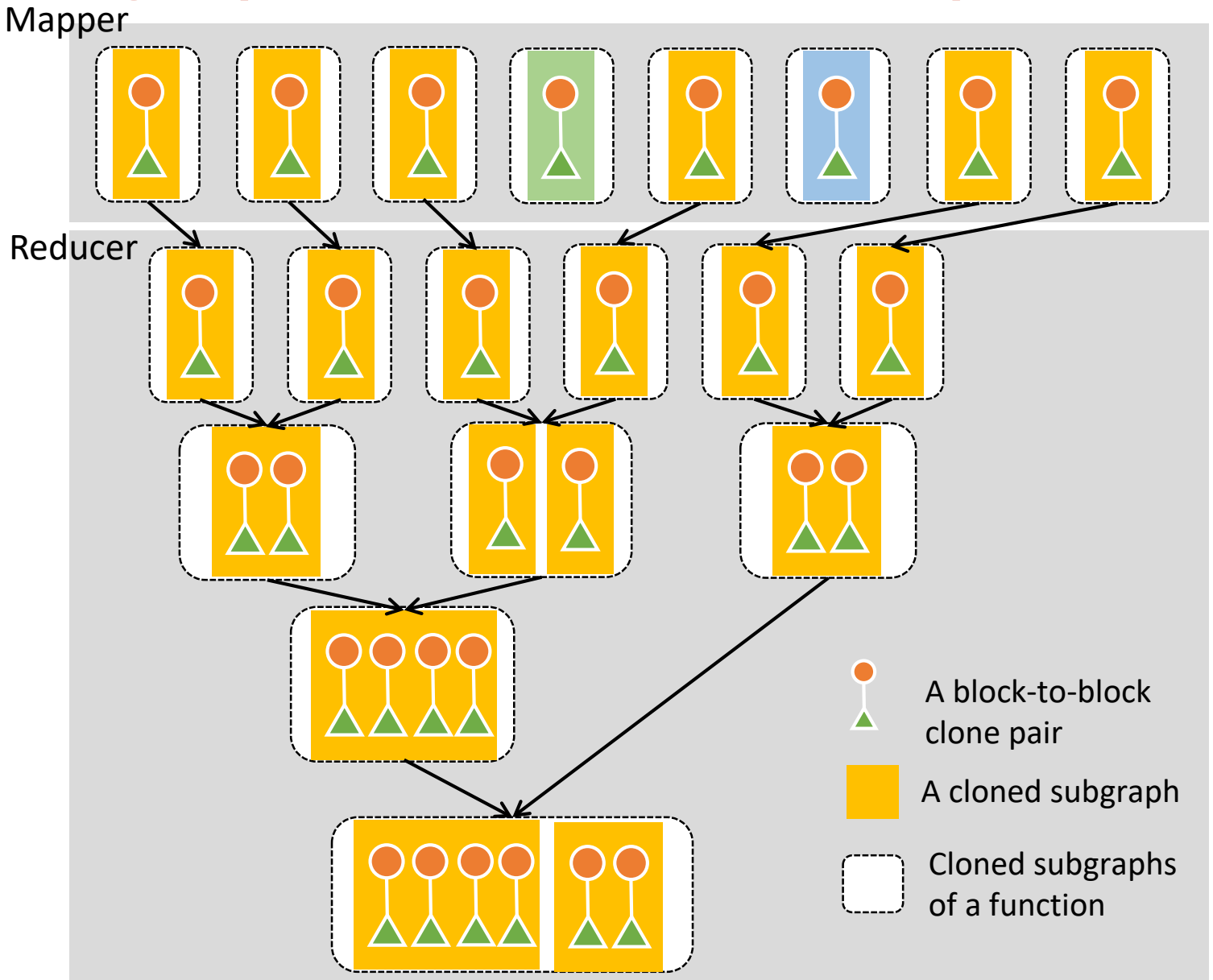
Efficiency (*k*-NN problem)

Adaptive Locality Sensitive Hashing

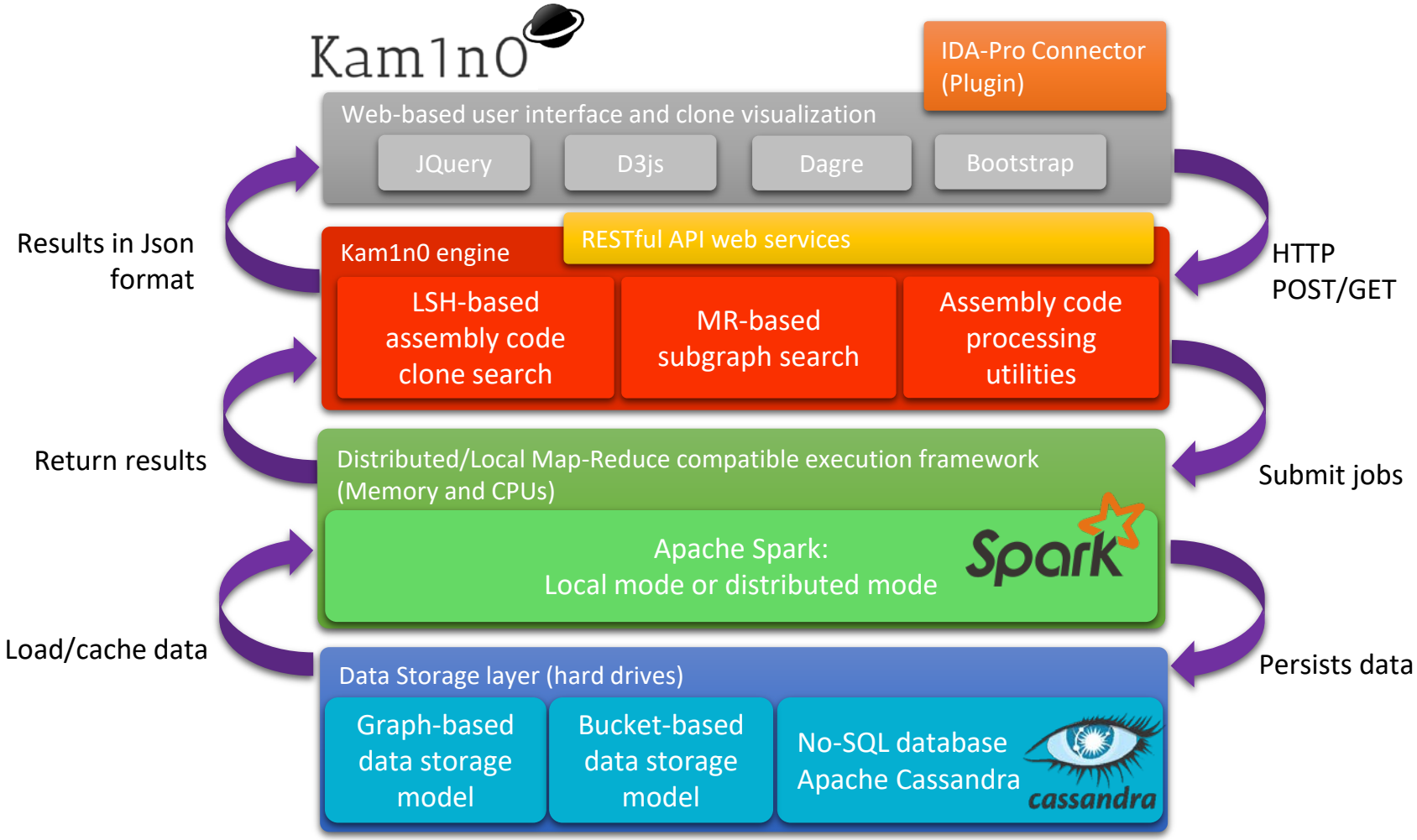
Unequal partitioning and dynamic hashing for cosine space



Subgraph Search with MapReduce



Overall Solution Stack



Published in ACM SIGKDD 2016

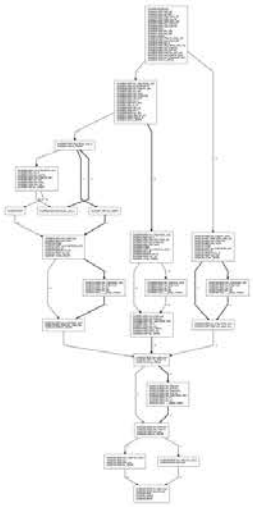
Software Demonstration: Kam1n0 ← Kamino

Evaluation (quality)

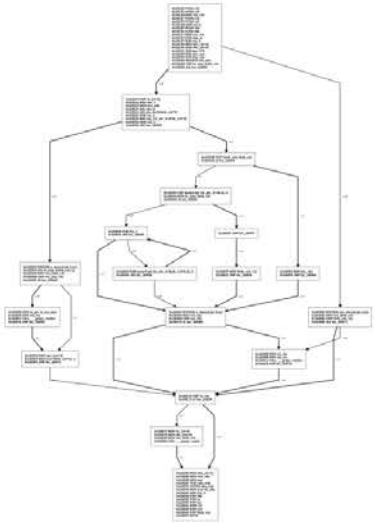
Approach	Bzip2	Curl	Expat	Jsoncpp	Libpng	Libtiff	Openssl	Sqlite	Tinyxml	Zlib	Avg.
BinClone	.985	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	*.894	.188
Composite	.857	.766	.693	.725	.814	.772	.688	.726	.688	.729	.746
Constant	.769	.759	.723	.665	.829	.764	.689	.776	.683	.768	.743
Graphlet	.775	.688	.673	.563	.714	.653	.682	.746	.676	.685	.685
Graphlet-C	.743	.761	.705	.604	.764	.729	.731	.748	.677	.668	.713
Graphlet-E	.523	.526	.505	.516	.519	.521	.512	.513	.524	.514	.517
MixGram	.900	.840	.728	.726	.830	.808	.809	.765	.707	.732	.785
MixGraph	.769	.733	.706	.587	.755	.692	.713	.765	.674	.708	.710
N-gram	.950	.860	.727	.713	.843	.809	.819	.789	.714	.766	.799
N-perm	.886	.847	.731	.729	.834	.813	.811	.769	.709	.736	.787
Tracelet	.830	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	.799	.163
LSH-S	.965	.901	.794	.854	.894	.922	.882	.845	.768	.758	.858
Kam1n0	*.992	*.989	*.843	*.890	*.944	*.967	*.891	*.895	*.864	.830	*.911

Area Under the Receiver Operating Characteristic Curve (AUROC)

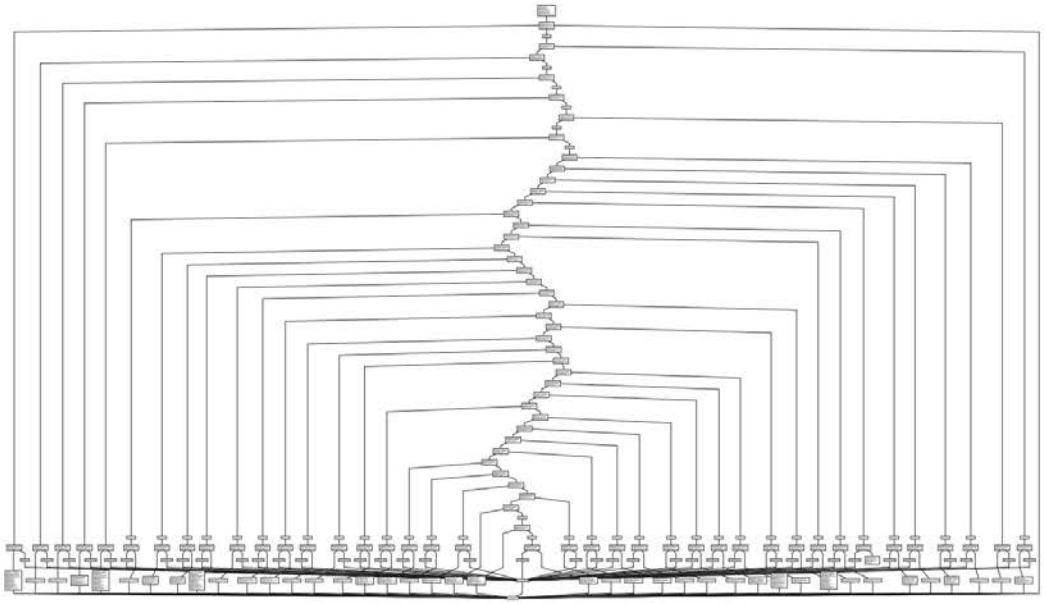
Obfuscation and Optimization - Challenges



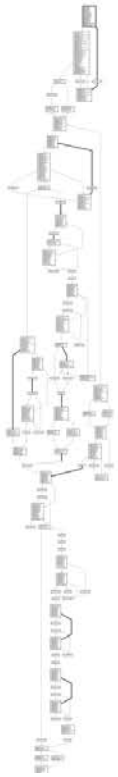
gcc O0



gcc O3



O-LLVM CFG Flattening



O-LLVM Bogus CFG

Acknowledgement: Current Students

NSERC CRSNG

Canada Research Chairs
www.chairs-chaires.gc.ca

DEFENCE R&D DÉFENSE

Fonds de recherche sur la nature et les technologies Québec

NCFTA
National Cyber-Forensics and Training Alliance CANADA

Doctoral Student	Status
Malik Altakrori	PhD candidate
Milad Ashouri	PhD candidate
Mohammad Reza Farhadi	PhD student
Rashid Hussain Khokhar	PhD candidate
Angel T. C. Lam	PhD student
Jwen-Fai Low	PhD candidate
Miles Q. Li	PhD candidate
+ 11 Master's students	

Acknowledgement: Former Students

Doctoral Student	Current Position
Khalil Al-Hussaeni	Assistant Professor, UAE
Sarah Alkhodair	Assistant Professor, Saudi
Rui Chen	Head of Data Science, Samsung Research America
Gaby Dagher	Assistant Professor, USA
Steven Ding	Assistant Professor, Canada
Farkhund Iqbal	Associate Professor, UAE
Noman Mohammed	Assistant Professor, Canada
Zhun Yu	Associate Professor, China
+ 19 Masters	

Thank you. Questions?

Benjamin Fung
McGill Data Mining and Security Lab
http://dmas.lab.mcgill.ca/fung
E-mail: ben.fung (at) mcgill.ca