Practical Evaluation of Static Analysis Tools for Cryptography: Benchmarking Method and Case of Study

Members:
- Tiago Moreira Trocoli da Cunha (226078)
- Eduardo Ito (159086)
- Albany Pinho (012832)
- Maria Júlia B. de Sousa (117964)
- Lucas André (182495)

Despite of availability of Static Code Analysis Tools (SCATs) only 35% of cryptographics misuses on such tools are detected.

--- Braga et al

Table of Contents

Objective

1) Compare different tools showing their limitations and strengths.

2) Determine how well tools perform in the context of cryptographic software development.

Contribution

1- A method for SCAT evaluation in detecting cryptography misuse

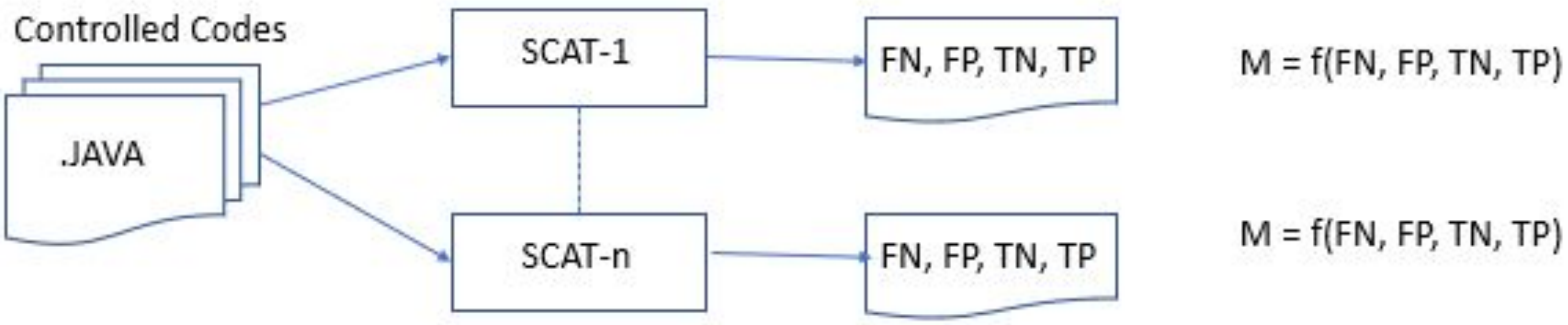
2- A set of realistic test cases for cryptographic misuse in Java

3- Assessment of free SCATs showing actual gaps in crypto misuse coverage

4- The evaluation of the tools according to methods defined hereby

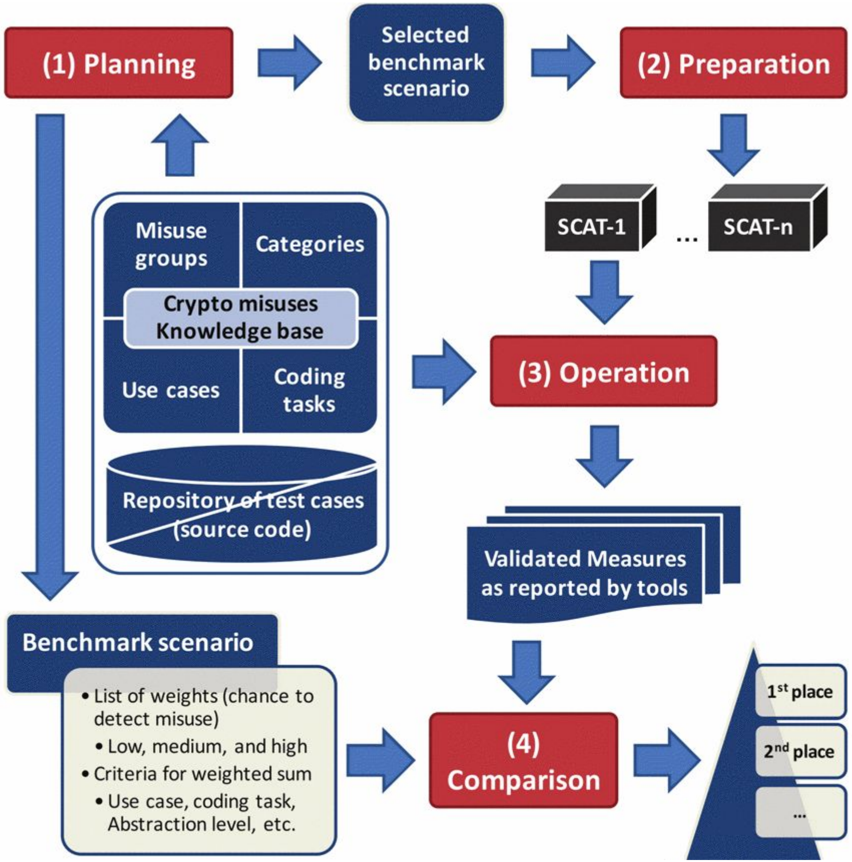
5- Recommendations for SCAT usage in specific development scenarios

MOTIVAÇÃO

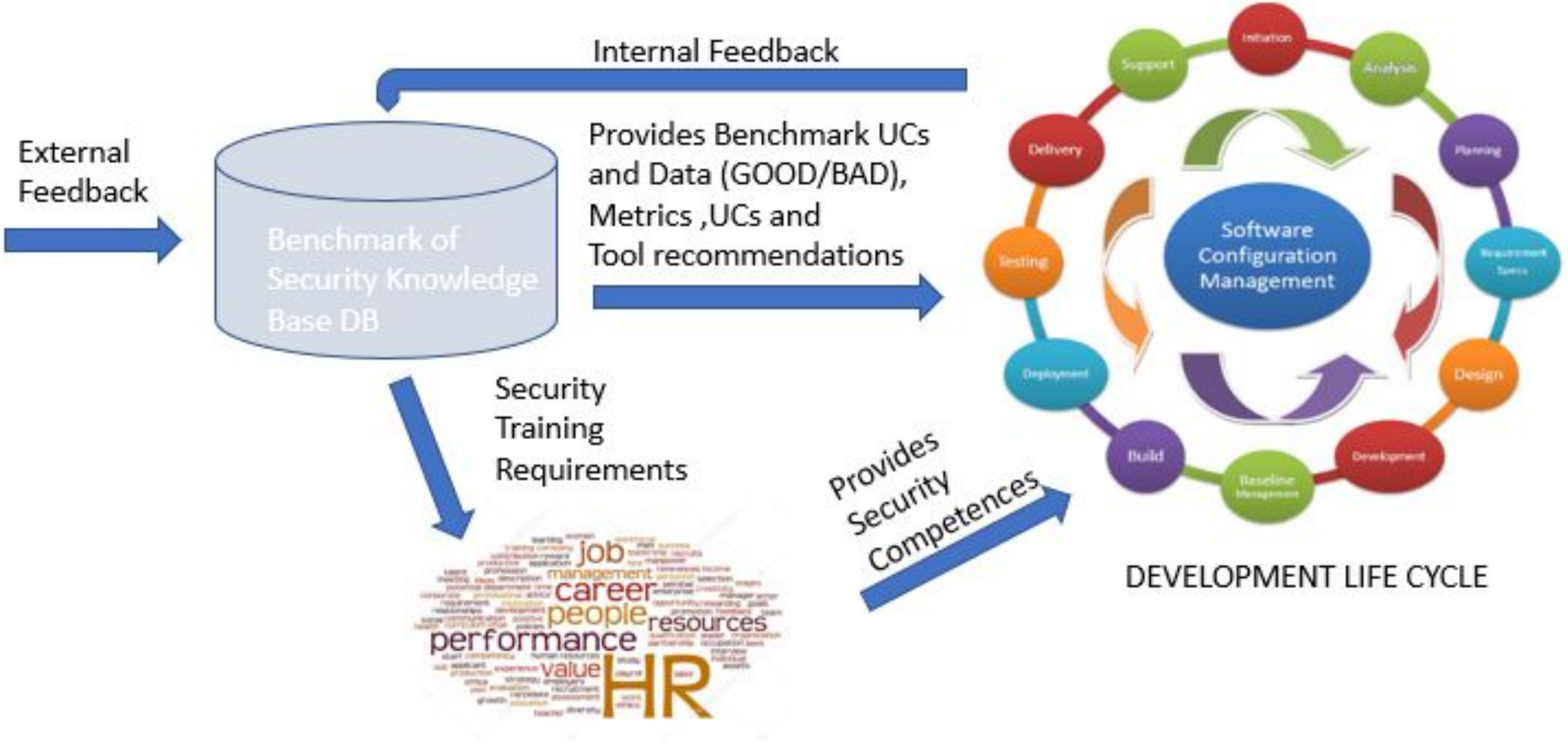


FP, TP = WC, CIB, BR, PDF, ICV, PKC, IVM, PKM, CAI(JAVA)

Methodology



RELATIONSHIP OF SECURITY BENCHMARK WITH DEVELOPMENT LIFE CYCLE



Low Complexity Category

MG	Misuse category	Misuse subtype
Misuse Group 1 (MG1)	Weak Cryptography (WC)	- Risky or broken crypto - Proprietary cryptography - Determin. symm. encryption - Risky or broken hash/MAC - Custom implementation
	Coding and Implementation Bugs (CIB)	- Wrong configs for PBE - Common coding errors - Buggy IV generation - Null cryptography - Leak/Print of keys
	Bad Randomness (BR)	- Use of statistic PRNGs - Predict., low entropy seeds - Static, fixed seeds - Reused seeds

Medium Complexity Category

Misuse Group 2 (MG2)	Program Design Flaws (PDF)	- Insecure default behavior - Insecure key handling - Insecure use of streamciphers - Insecure combo encrypt/auth - Insecure combo encrypt/hash - Side-channel attacks
	Improper Certificate Validation (ICV)	- Missing validation of certs - Broken SSL/TLS channel - Incomplete cert. validation - Improper validated host/user - Wildcards, self-signed certs
	Public-Key Cryptography (PKC) issues	- Deterministic encrypt. RSA - Insecure padding RSA enc. - Weak configs for RSA enc. - Insecure padding RSA sign. - Weak signatures for RSA - Weak signatures for ECDSA - Key agreement: DH/ECDH - ECC: insecure curves

High Complexity Category

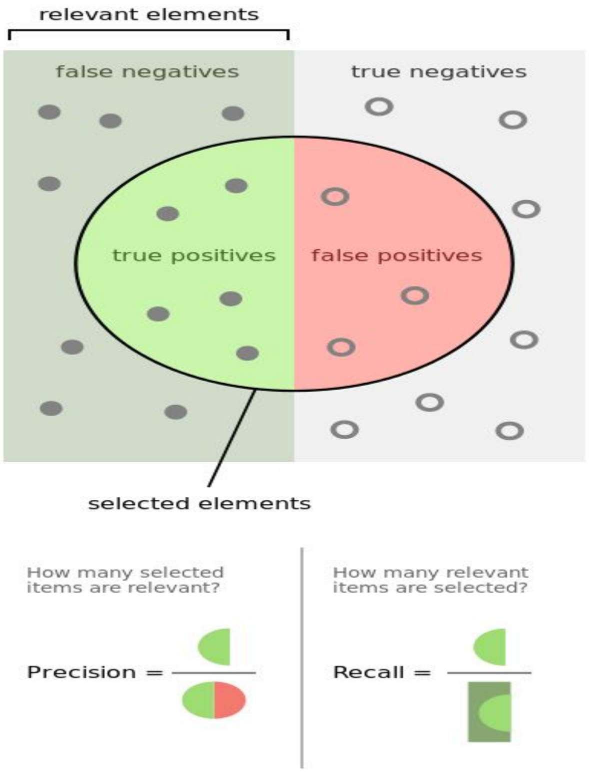
Misuse Group 3 (MG3)	IV and Nonce Management (IVM) issues	- CBC with non-random IV - CTR with static counter - Hard-coded or constant IV - Reused nonce in encryption
	Poor Key Management (PKM)	- Short key, improper key size - Hard-coded or constant keys - Hard-coded PBE passwords - Reused keys in streamciphers - Use of expired keys - Key distribution issues
	Crypto Architecture and Infrastructure (CAI) issues	- Crypto agility issues - API misunderstanding - Multiple access points - Randomness source issues - PKI and CA issues

Basic Concept: Confusion Matrix

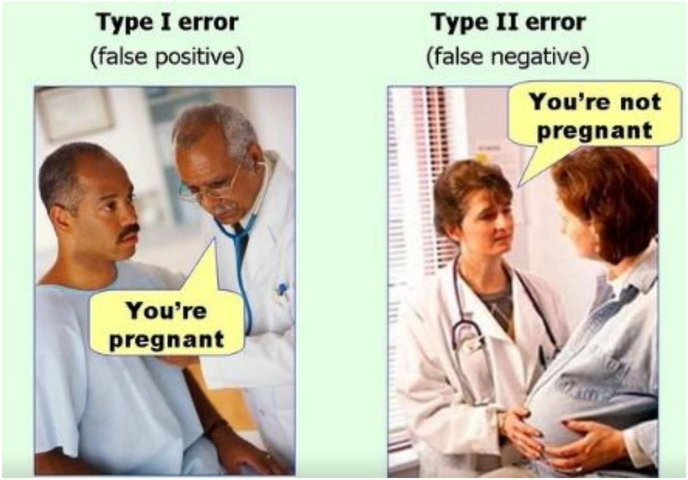
		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

https://en.wikipedia.org/wiki/Precision_and_recall

Basic concept: Precision and Recall



- POSITIVE: cryptographic misuse, vulnerability found
- NEGATIVE: correct use of cryptography, no vulnerability



Example

- 22 test cases which 10 are misuses and 12 are good uses.
- A SCAT reported 13 reported as positive cases which 8 are true misuses and 5 are actually false alarms.

Test Case	**Reported by SCAT**	
	Reported Positive (P)	**Reported Negative (N)**
Positive (misuse)	8 (TP)	2 (FN)
Negative (good use)	5 (FP)	7 (TN)

Oracle Test Case	Reported by Evaluated Tool	
	Reported Positive	Reported Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Metrics	
Recall	TP / (TP + FN)
Precision	TP / (TP + FP)
F-Measure	2TP / (2TP + FN + FP)

TABLE III

CONTEXTS COMBINING TEAMS AND APPLICATION PROFILES.

C#	**Context**	**App profile**	**Misuse groups**
`C1`	Novice team, no expert	Low complexity	MG1
`C2`	Novice team with expert	Low to medium	MG1 and MG2
`C3`	Skilled team, no expert	Medium to high	MG2 and MG3
`C4`	Skilled team and expert	High complexity	MG3

1. **Unsupported Novice Team (C1)**
   - Team wants to avoid calling out experts for false alarms (low FP) → Precision.
   - Low false negative (FN) helps optimize time for experts → Recall.
2. **Supported Novice Team (C2)**
   - Team wants solve as much as true positive cases with experts (low FN)→Recall.
   - Low false alarms helps experts → Precision.
3. **Unsupported Knowledgeable Team (C3)**
   - Experienced developers want to solve as many of the issue, in a best effort approach → F-measure
   - In complex cases, experienced developers may not detect all omission cases, that experts can solve, but support not always available (Low FN) → Recall
4. **Supported Knowledgeable Team (C4)**
   - Team members can easily identify FP, so the most important is FN → Recall
   - Low FP is time-saving → Precision

Based on our previous observations, we can create a table of context x priority metrics.

TABLE IV
DEVELOPMENT CONTEXTS AND RECOMMENDED METRICS.

Context	**1st. Metric**	**2nd. Metric**
C1	Precision	Recall
C2	Recall	Precision
C3	F-Measure	Recall
C4	Recall	Precision

Scenarios

TABLE V
WEIGHTS FOR SCENARIOS, CONTEXT, AND MISUSE GROUPS.

Scenario	Context	Weights per misuse group		
		MG1	MG2	MG3
S1	C1	High	Low	Low
S2	C2	Medium	High	Low
S3	C3	Low	Medium	High
S4	C4	Low	Low	High

- 220 misuses (+)
- 182 good uses (-)

- 384 test programs (Java)

TABLE VI
CRYPTO MISUSES DISTRIBUTED BY CATEGORIES.

Criteria	Subset	Misuse	Good use
Misuse group	MG1	61	34
	MG2	106	99
	MG3	35	49
Crypto use cases	EDR	90	93
	AVD	49	39
	RND	13	10
	PPE	10	5
	SC	40	35
Crypto coding tasks	Enc/Dec	83	68
	Sign/Ver	26	27
	Hash/MAC	22	11
	KG	43	61
	SSL	10	3
	Cert	5	2
	Rand	12	10
Misuse categories	WC	20	10
	CIB	29	16
	BR	12	8
	PDF	23	14
	ICV	15	5
	PKC/ENC	27	30
	PKC/SIG	21	25
	PKC/ECC	14	20
	PKC/KA	6	5
	IVM	8	10
	PKM	19	32
	CAI	8	7

SCATS

FindSecBug 1.5.0 (FSB)

VisualCodeGrepper 2.1.0 (VCG)

Xanitizer (Xan)

SonarQube (SQ)

Yasca

Results

TABLE VII
RESULTS FOR FIVE FREE SCATS.

Tools	Measures				Metrics		
	TP	TN	FN	FP	Prec.	Recall	F-M
FSB	51	147	151	35	0.593	0.252	0.354
Xan	68	140	134	42	0.618	0.337	0.436
SQ	5	181	197	1	0.833	0.025	0.048
VCG	7	180	195	2	0.778	0.035	0.066
Yasca	8	182	194	0	1.000	0.040	0.076

- Xan detected ⅓ of all misuses.
- All tools detected only 35% of all misuses
- Xan has the highest TPs and lowest FNs → better tool.

Results

TABLE VIII
RESULTS FOR MISUSE GROUP ONE (MG1).

Tools	Metrics for WC			Metrics for CIB			Metrics for BR		
	Prec.	Recall	F-M	Prec.	Recall	F-M	Prec.	Recall	F-M
FSB	0.727	0.40	0.516	0.50	0.172	0.256	1.0	0.417	0.588
Xan	0.588	0.50	0.541	0.70	0.483	0.571	1.0	0.417	0.588
SQ	1.0	0.20	0.333	0.0	0.0	0.0	0.0	0.0	0.0
VCG	1.0	0.20	0.333	0.0	0.0	0.0	1.0	0.250	0.400
Yasca	1.0	0.30	0.462	0.0	0.0	0.0	1.0	0.167	0.286

Weak Cryptography (WC):

- Xan and FSB are the best (high recall).
- Most tools detected DES and 3DES.
- Yasca didn’t detected SHA-1.
- Only Xan detected weak hash functions.
- No tool detected BlowFish, RC4 and insecure PBE.

Results

TABLE VIII
RESULTS FOR MISUSE GROUP ONE (MG1).

Tools	Metrics for WC			Metrics for CIB			Metrics for BR		
	Prec.	Recall	F-M	Prec.	Recall	F-M	Prec.	Recall	F-M
FSB	0.727	0.40	0.516	0.50	0.172	0.256	1.0	0.417	0.588
Xan	0.588	0.50	0.541	0.70	0.483	0.571	1.0	0.417	0.588
SQ	1.0	0.20	0.333	0.0	0.0	0.0	0.0	0.0	0.0
VCG	1.0	0.20	0.333	0.0	0.0	0.0	1.0	0.250	0.400
Yasca	1.0	0.30	0.462	0.0	0.0	0.0	1.0	0.167	0.286

Code Implementation (CIB):

- SQ, VCQ and Yasca didn't score TP and FP.
- Xan had the most TP but high FP (false alarms) -> High Precision.
- Xan and FSB detected Buggy IV and NullCypher.
- Only Xan detected leak of privacy.
- None detected saved keys in strings.

Results

TABLE VIII
RESULTS FOR MISUSE GROUP ONE (MG1).

Tools	Metrics for WC			Metrics for CIB			Metrics for BR		
	Prec.	Recall	F-M	Prec.	Recall	F-M	Prec.	Recall	F-M
FSB	0.727	0.40	0.516	0.50	0.172	0.256	1.0	0.417	0.588
Xan	0.588	0.50	0.541	0.70	0.483	0.571	1.0	0.417	0.588
SQ	1.0	0.20	0.333	0.0	0.0	0.0	0.0	0.0	0.0
VCG	1.0	0.20	0.333	0.0	0.0	0.0	1.0	0.250	0.400
Yasca	1.0	0.30	0.462	0.0	0.0	0.0	1.0	0.167	0.286

Bad Randomness (BR):

- No tool detected fixed seeds nor reuse.
- SQ didn’t score.
- Xan and FSB detected all statics PRNG.
- VCG and Yasca got low recall.

Results

TABLE IX
RESULTS FOR MISUSE GROUP TWO (MG2).

Tools	Metrics for PDF			Metrics for PKC			Metrics for ICV		
	Prec.	Recall	F-M	Prec.	Recall	F-M	Prec.	Recall	F-M
FSB	0.417	0.217	0.286	1.0	0.221	0.361	1.0	0.267	0.421
Xan	0.357	0.217	0.270	1.0	0.235	0.381	1.0	0.133	0.235
SQ	0.0	0.0	0.0	1.0	0.015	0.029	0.0	0.0	0.0

Program Design Flow (PDF):

- SQ, VCG and Yasca didn't score.
- FSB -> highest Prec.
- FSB can detected AES insecure default.
- Other insecure default (RSA, PBE, OAEP...) were not detected.
- Tool didn't score well (blind spots).

Results

TABLE IX
RESULTS FOR MISUSE GROUP TWO (MG2).

Tools	Metrics for PDF			Metrics for PKC			Metrics for ICV		
	Prec.	Recall	F-M	Prec.	Recall	F-M	Prec.	Recall	F-M
FSB	0.417	0.217	0.286	1.0	0.221	0.361	1.0	0.267	0.421
Xan	0.357	0.217	0.270	1.0	0.235	0.381	1.0	0.133	0.235
SQ	0.0	0.0	0.0	1.0	0.015	0.029	0.0	0.0	0.0

Public-Key Cryptography (PKC):

- VCG and Yasca didn’t score.
- All others → highest precision (no FP).
- Xan wins → highest recall and f-measure.
- No one detected insecure hash.
- SQ bug → case sensitive for algorithm names.

Results

TABLE IX
RESULTS FOR MISUSE GROUP TWO (MG2).

Tools	Metrics for PDF			Metrics for PKC			Metrics for ICV		
	Prec.	Recall	F-M	Prec.	Recall	F-M	Prec.	Recall	F-M
FSB	0.417	0.217	0.286	1.0	0.221	0.361	1.0	0.267	0.421
Xan	0.357	0.217	0.270	1.0	0.235	0.381	1.0	0.133	0.235
SQ	0.0	0.0	0.0	1.0	0.015	0.029	0.0	0.0	0.0

Improper Certificate Validation (ICV):

- FSB and Xan detected certificate validation related to SSL/TLS.
- Xan has a bug that prevented the detection of misuse in nested classes.
- FSB got better recall and f-measure due to higher TP.

Results

TABLE X
RESULTS FOR MISUSE GROUP THREE (MG3).

Tools	Metrics for IVM			Metrics for PKM			Metrics for CAI		
	Prec.	Recall	F-M	Prec.	Recall	F-M	Prec.	Recall	F-M
FSB	0.286	0.500	0.364	0.263	0.263	0.263	0.0	0.0	0.0
Xan	0.231	0.375	0.286	0.263	0.263	0.263	0.800	1.0	0.889

IV and Nonce Management:

- VCG, SQ and Yasca didn't score.
- No tools detected non-random IV for CTR.
- No tools detected static counter for CTR.
- High FP → tools had difficulties to understand program design for IV management.

Results

TABLE X
RESULTS FOR MISUSE GROUP THREE (MG3).

Tools	Metrics for IVM			Metrics for PKM			Metrics for CAI		
	Prec.	Recall	F-M	Prec.	Recall	F-M	Prec.	Recall	F-M
FSB	0.286	0.500	0.364	0.263	0.263	0.263	0.0	0.0	0.0
Xan	0.231	0.375	0.286	0.263	0.263	0.263	0.800	1.0	0.889

Poor Key Management (PKM):

- VCG, SQ and Yasca did not score.
- FSB and Xan were tied.

Results

TABLE X
RESULTS FOR MISUSE GROUP THREE (MG3).

Tools	Metrics for IVM			Metrics for PKM			Metrics for CAI		
	Prec.	Recall	F-M	Prec.	Recall	F-M	Prec.	Recall	F-M
FSB	0.286	0.500	0.364	0.263	0.263	0.263	0.0	0.0	0.0
Xan	0.231	0.375	0.286	0.263	0.263	0.263	0.800	1.0	0.889

Crypto Architecture and Infrastructural (CAI):

- Most difficult to detect.
- But Xan obtained good results.

Results

TABLE XI
WEIGHTED METRICS FOR FIVE SCATS IN FOUR SCENARIOS.

Tools	Weighted metrics for S1			Weighted metrics for S2			Weighted metrics for S3			Weighted metrics for S4		
	W-Prec.	W-Recall	W-F-M	W-Prec.	W-Recall	W-F-M	W-Prec.	W-Recall	W-F-M	W-Prec.	W-Recall	W-F-M
Xan	**0.737**	0.451	0.537	0.756	**0.302**	0.392	0.563	0.431	**0.427**	0.488	**0.510**	0.471
FSB	0.701	0.316	0.425	0.747	0.266	0.377	0.412	0.253	0.270	0.281	0.259	0.242
Yasca	0.556	**0.130**	0.208	0.208	**0.049**	0.078	0.042	0.010	**0.016**	0.056	0.013	0.021
VCG	0.556	0.125	0.204	0.208	0.047	0.076	0.042	0.009	0.015	0.056	0.013	0.020
SQ	0.306	0.056	0.093	0.313	0.024	0.041	0.125	0.006	0.010	0.056	0.006	0.010

Conclusion

About the tools

- Benchmark of cryptographic misuse helps to categorize static analysis tools (SCATs)
- Recommended metrics (precision, recall, and f-measure)

Secure software development

- Useful for scheduling activities during project planning
- Dimension resources based on context and complexity of applications
  - Based on developer skills and expert availability

Conclusion

- Tests based on Java
- Focus on cryptography misuses, may perform better with other security domains
- Newer versions of tools may change (improve?) results
- Free tools recommended for MG1 cases, or MG2 with expert help, not suitable for advanced scenarios

TABLE XIII
CONTEXTS LINK LIKELY MISUSES AND TOOL USAGE.

Context	Misuse group	Usage	Tool
C1	MG1	Integrated to IDE	Xan and FSB
C2	MG1 and MG2	IDE and build	Xan and FSB
C3	MG2 and MG3	Build and review	None
C4	MG3	Reviews	None

Questions???

References

[1] A. Braga and R. Dahab. Practical evaluation of static analysis tools for cryptography: benchmarking method and case study. *IEEE 28th International Symposium on Software Reliability Engineering*, Oct. 2017.

[2] C. Paar and J. Pelzl. Understanding cryptography. *ACM Computing Classification (1998): E.3, K.4.4, K.6.5.*, 1998.

Acronyms

PDF	Program design flaws
PKC	Public-key cryptography
PKM	Poor Key Management
PPE	Password Protection with Encryption
PRNG	Pseudorandom number generator
SC	Secure communication
WC	Weak cryptography

Acronyms

AVD	Authentication and Validation of Data
BR	Bad Randomness
CAI	Crypto Architecture and Infrastructure
CIB	Coding and implementation bugs
ICV	Improper certificate validation
EDR	Encrypt Data at Rest
IVM	IV and nonce management