AWSomePy: A Dataset and Characterization of Serverless Applications

G. Raffa, J. Blasco, D. O'Keeffe, S. K. Dash
Royal Holloway, University of London
EuroSys 2023 Conference
SESAME '23 Workshop, 8th May 2023

ROYAL HOLLOWAY UNIVERSITY Of LONDON

Introduction

- **Serverless paradigm challenges**
  - Performance
  - Traceability
  - Security

- **Static and dynamic analysis**
  - Variety of sources and events
  - Existing analysis frameworks not optimized
  - Models / approximations needed for static analysis

Development of new models and tools → **Characterization of real-world applications**

Research Objective & Outline

- **Objective**
  - Identification of key trends in serverless applications

AWSomePy Dataset Generation	→	**145 AWS Applications Implemented in Python**
Configuration & Architectural Analysis	→	**Plugins, Lines of Code & No. of Handlers / Events**
Application Code-level Analysis	→	**Cloud Platform Services & API Usage**

Dataset Generation (1)

- **Wonderless dataset** (*)
  - Multiple languages / providers
  - No metadata (e.g., no. of handlers)
  - Created in July 2020

- **AWSomePy dataset**
  - AWS and Python focused
  - New processing step
  - Created in August 2022

Repositories Filtering by Language & Cloning

Metadata Gathering

(*) N. Eskandani and G. Salvaneschi, "The Wonderless Dataset for Serverless Computing," *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*, Madrid, Spain.

Dataset Generation (2)

- **AWSomePy metadata**
  - Publicly released in a CSV file

LOC	Stars	Watching	Forks	SF_Version	Num_Of_Plugins
10226	54	7	18	N/A	1
1297	1	1	0	>=2.0.0<3.0.0	1
190	3	2	0	N/A	0
30	0	2	0	N/A	1
513	5	1	3	N/A	1
370	4	2	1	N/A	1

Repository-specific Metadata → **No. of Stars, Watchers & Forks**

Application-level Metadata → **SF Version, LOC, No. of Events / Handlers**

Config. & Architectural Analysis (1)

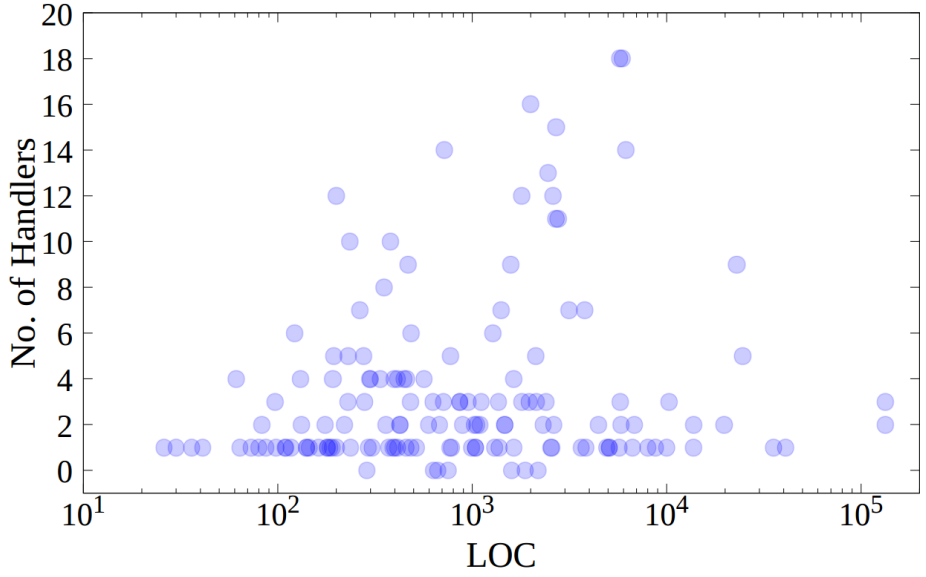
- **Plugin analysis**
  - Specified in infrastructure code file (YAML)
  - 44 plugins in total
- **Results**
  - 1st & 2nd => configuration
  - 3rd & 4th => functionality

Plugins	Occurrences
serverless-python-requirements	95
serverless-pseudo-parameters	25
serverless-domain-manager	15
serverless-step-functions	14
serverless-offline	9
serverless-dotenv-plugin	8
serverless-prune-plugin	8
serverless-iam-roles-per-function	7

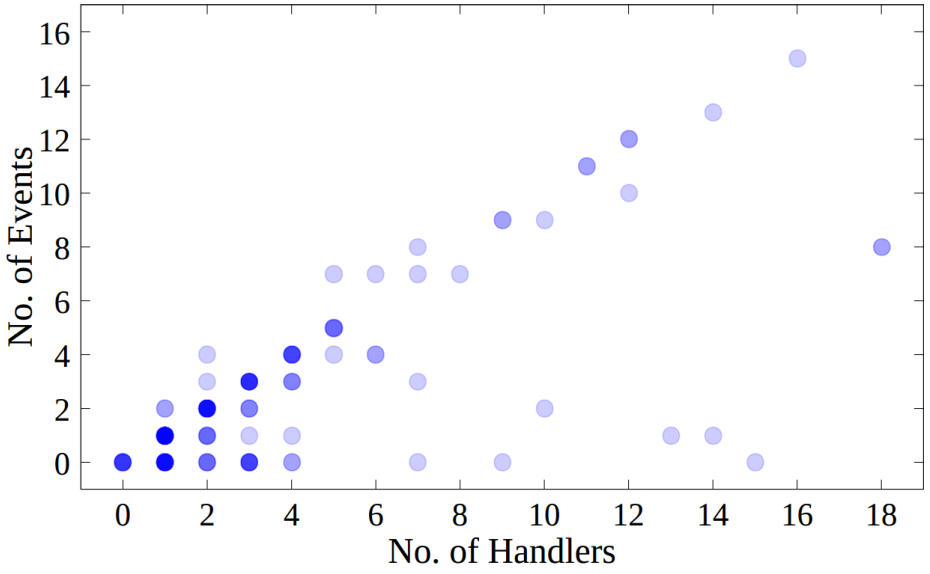
Developers are not configuring permissions in a granular fashion

Config. & Architectural Analysis (2)

- **Complexity analysis**
  - Lines of Code (LOC)
    - ≈ 45% of the repositories 100 < LOC < 1k
  - Infrastructure code processing



Typically ≤ 4 handlers



Majority ≤ 5 events

Application Code-level Analysis (1)

- **Cloud services**
  - Client or resource objects

boto3.client('s3')

boto3.resource('s3')

  - 46 services in total

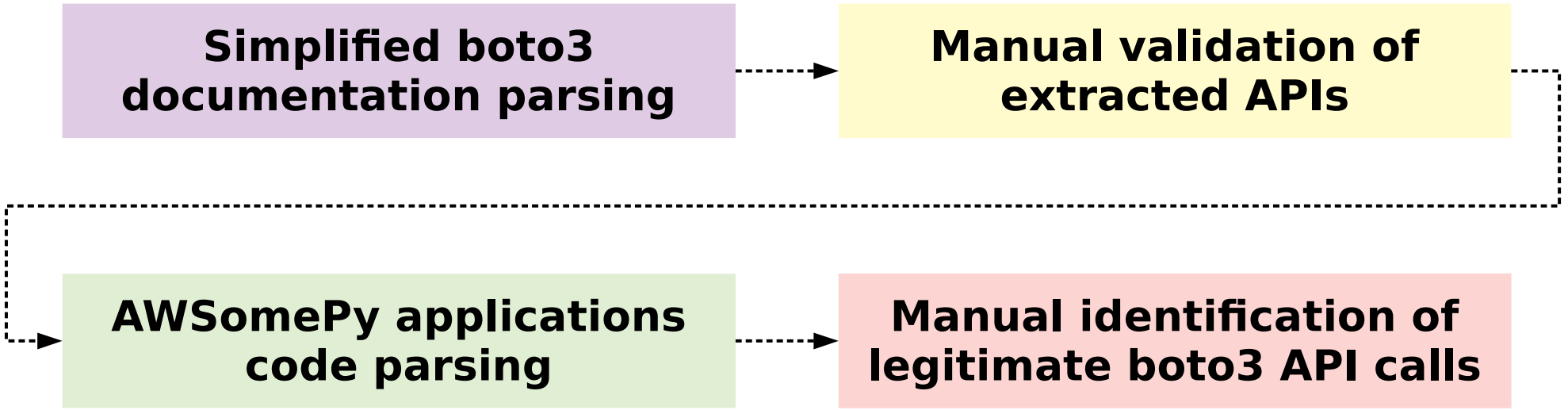
Services	No. of Repositories	Occurrences
s3	59	217
dynamodb	47	201
lambda	24	47
ssm	14	46
sqs	21	41
sns	11	30
ec2	12	29
sts	9	26
rekognition	8	15
cloudformation	7	14
stepfunctions	9	14

Data storage and NoSQL services the most common

Configuration-oriented services frequently used

Application Code-level Analysis (2)

- **Cloud APIs**
  - Semi-automated approach
  - Accuracy / ease of implementation



Application Code-level Analysis (3)

- **Cloud APIs**
  - Programmatic creation of buckets & tables
    - Resources cannot be checked via infrastructure code analysis
  - Use of invoke API to trigger handler execution
    - Added workflows not easily detectable via static analysis

s3 API	#	**dynamodb** API	#	**lambda** API	#
put_object	61	put_item	143	invoke	55
get_object	52	scan	64	add_permission	7
create_bucket	50	query	62	list_functions	3
upload_file	48	get_item	58	get_policy	3
download_file	24	update_item	57	get_function	2
list_objects_v2	22	create_table	41	list_tags	2
other	111	*other*	93	*other*	4

Conclusion

- **Key takeaways**
  - All security-related

Granular configuration of handler permissions	→	**Not widely adopted in AWSomePy**
Configuration and management services	→	**Workflows difficult to inspect before deployment**
Programmatic creation of data stores & tables	→	**Resources cannot be checked before deployment**

Thank You!

- **AWSomePy dataset**

https://doi.org/10.5281/zenodo.7838076

- **Any questions?**

Email: **giuseppe.raffa.2018@live.rhul.ac.uk**