SDS2023
IEEE SWISS CONFERENCE ON DATA SCIENCE

The 10th IEEE Swiss Conference on Data Science
June 22 – 23, 2023, Zürich, Switzerland

IEEE

A Graph-Representation-Learning Framework for Supporting Android Malware Identification and Polymorphic Evolution

Alfredo Cuzzocrea, Miguel Querbado, Abderraouf Hafsaoui, Edoardo Serra

iDEA Lab, University of Calabria, Rende, Italy

Computer Science Dept., Boise State University, Boise ID, USA

Work Plan

1. Introduction
2. Contextual Knowledge
3. Proposed Approach
4. Experimental Evaluation
5. Conclusion & Future Work

Introduction

- **Cybersecurity** is an interesting research area that has emerged in many scientific fields.
- **Cybersecurity** plays a central role due to the increasing reliance on digital technologies and the interconnectedness of systems.
- The consequences of **cyberattacks** can range from financial loss and reputational damage to compromise of sensitive information and disruption of critical infrastructure.

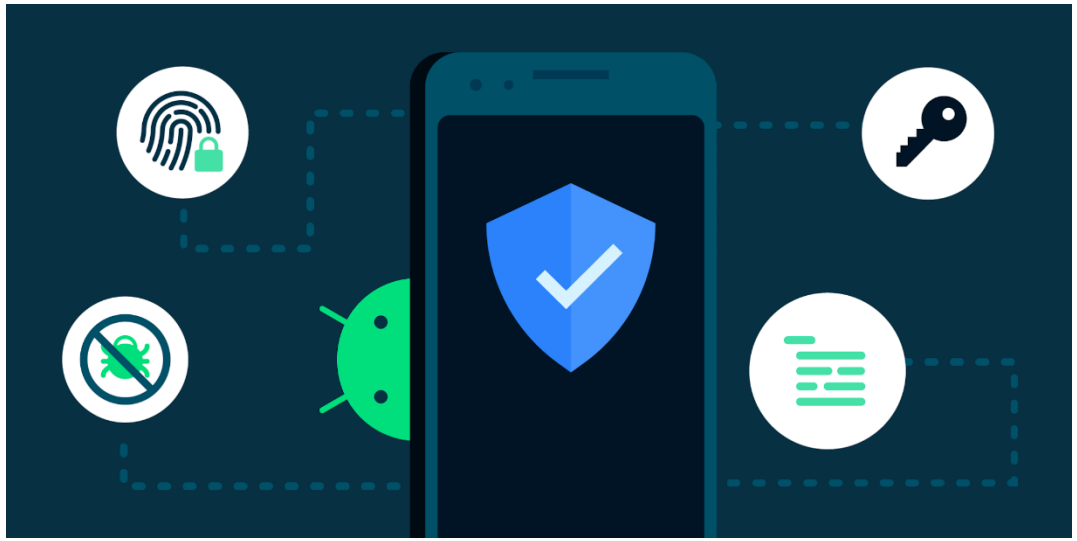
Introduction

- Continuous advancements in technology and evolving threat landscape necessitate ongoing research, and innovation in cybersecurity.

- Cybersecurity involves a combination of preventive measures, detection, and monitoring techniques.

Introduction

- Android cybersecurity focuses on the protection of the Android OS and devices from various cyberattacks.
- In this research, we focused our attention on Android cybersecurity, particularly **Malware Detection** in Android applications.



Contextual Knowledge

- **Graph-Representation-Learning** techniques have been adopted by various scientific domains for several essential downstream tasks.
- ***Inferential Structural Iterative Representation Learning Approach* (*Inferential SIR-GN*)** is used in this work in order to identify malware and classify *Android APK types and families*.
- This technique is applied to **MALNET-TINY**, a public dataset of Android APK files.

Contextual Knowledge

- **MALNET-TINY** is one of the biggest cybersecurity datasets that has been released, it involves more than 80K Android APK files.

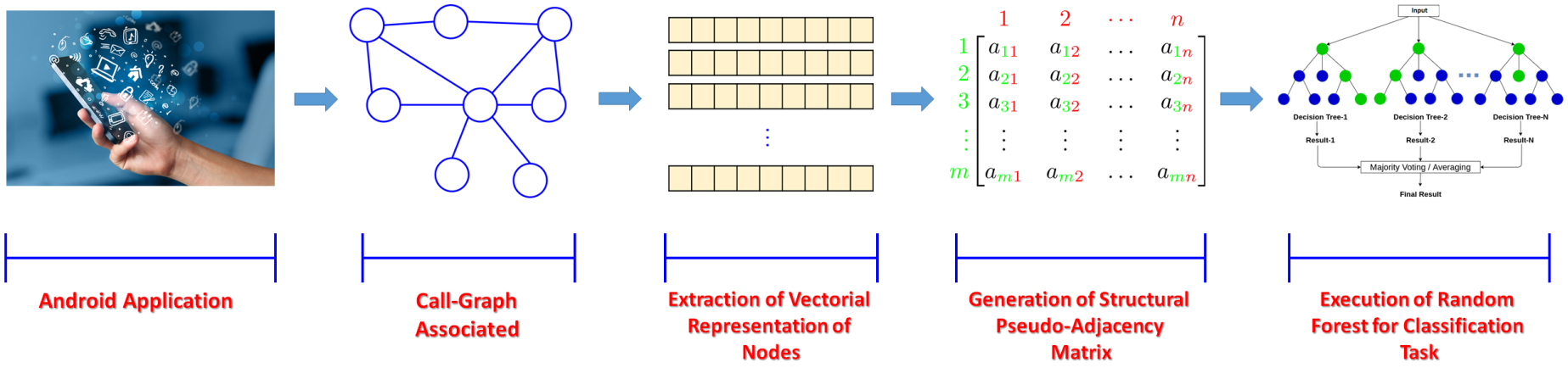
- **MALNET-TINY** enable researchers to prototype new concepts and ideas by providing a large-scale dataset to train new models.

Proposed Approach

- The process of developing a malware classifier that is resistant to malware polymorphism is divided into four steps:

  - ✓ Extraction of the ***Structural Vectorial Representation*** from the Call-Graph associated to the Android application.
  - ✓ Generation of the ***Structural Pseudo-Adjacency Matrix***.
  - ✓ Build a potentially ***Polymorphic Variant*** of the malware.
  - ✓ Training of the ***Random Forest*** to identify and classify malware.

Proposed Approach



Android Application

Call-Graph Associated

Extraction of Vectorial Representation of Nodes

Generation of Structural Pseudo-Adjacency Matrix

Execution of Random Forest for Classification Task

Proposed Approach

Extraction of the Structural Vectorial Representation

- **Inferential SIR-GN** technique is used to extract node representations from directed graphs, which characterizes and aggregates a node's neighbors to iteratively update its representation.
- The size of a node's representation at each iteration is equal to a user-specified hyperparameter $\boldsymbol{N}$.
- The current node representation, which starts with the node degree, is grouped into $\boldsymbol{N}$ KMeans clusters to create the node descriptions.

Proposed Approach

Extraction of the Structural Vectorial Representation

- These descriptions are then concatenated into a bigger representation that reflects the evolution of structural information via deeper neighborhood exploration.
- A ***Principle Component Analysis*** **(PCA)** is employed to reduce the size of the final node representations while preserving their most important features.
- Overall, this process allows the extraction of a structural vectorial representation for each node that captures both local and global structural information.

Proposed Approach

Generation of the Structural Pseudo-Adjacency Matrix

For each edge (u, v) in E, calculate the distance r_{ui} between node u and the centroid of its cluster i, and the distance r_{uj} between node v and the centroid of its cluster j

$$M_{(i,j)} = \sum_{(u,v)\in E} r_{ui} \times r_{vj}$$

Algorithm 1 `PseudoAdjacencyMatrix`

Input: Graph G, Matrix size w, Set of nodes V
Output: Structural Pseudo Adjacency Matrix $GraphRep$

Begin
  Initialize a matrix $GraphRep \in Z^{wXw}$ of zeros;
  for all $(u \in V)$ **do**
    for all $(nbr(u))$ **do**
      $GraphRep+ = Emb[u].reshape(1, w) * Emb[u].reshape(w, 1)$
    end for
  end for
  return $GraphRep$;
End

It takes as input a graph G with nodes V and edges E

Linearize the result matrix to form a vector of features

Proposed Approach

Build a Potentially Polymorphic Variant of the Malware

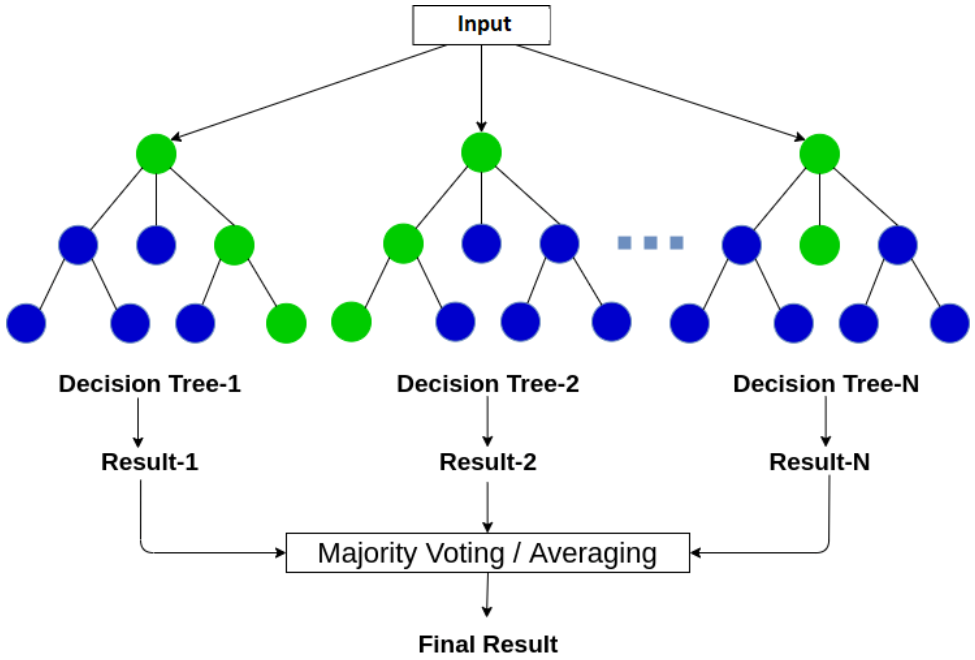
- Identify a benign Android application that has a Structural Pseudo-Adjacency Matrix (SPAM) closest in terms of Euclidean distance to that of the malware using the ***K-Nearest Neighbor*** technique (**KNN**).

- Combine the Structural Pseudo-Adjacency Matrix of the malware with that of the benign application to create a potentially polymorphic variant of the malware.

Proposed Approach

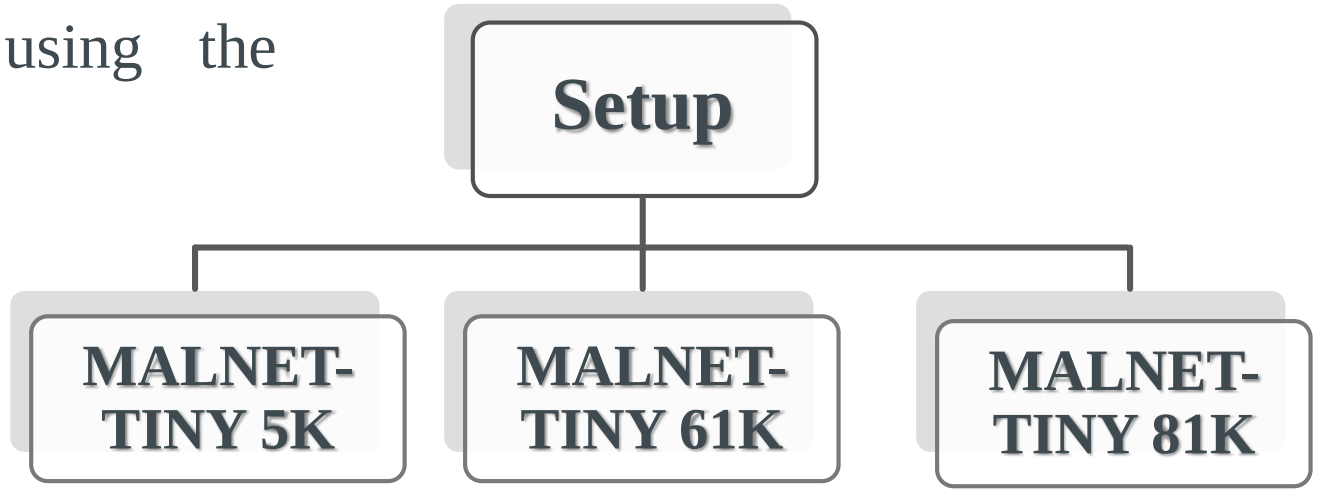
Training of the Random Forest for Identification and Classification

- The **Random Forest** model is trained on the set of features produced by the linearized matrix and the polymorphic representation in order to make it resistant to polymorphic changes in malware applications.



Experimental Evaluation

- The experimental assessments accomplished in order to evaluate the effectiveness of the proposed framework involve using the following three datasets.



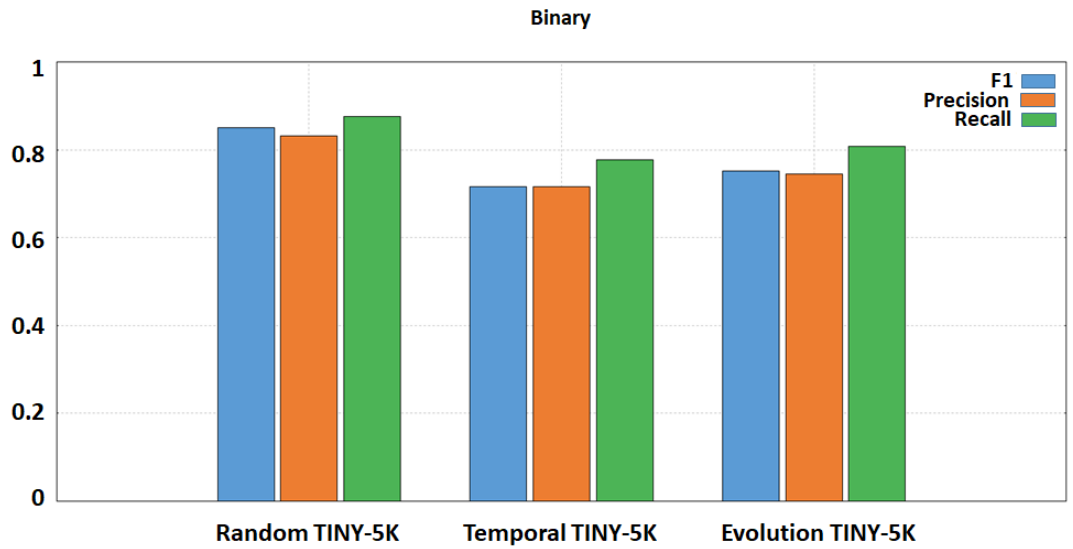
Experimental Evaluation

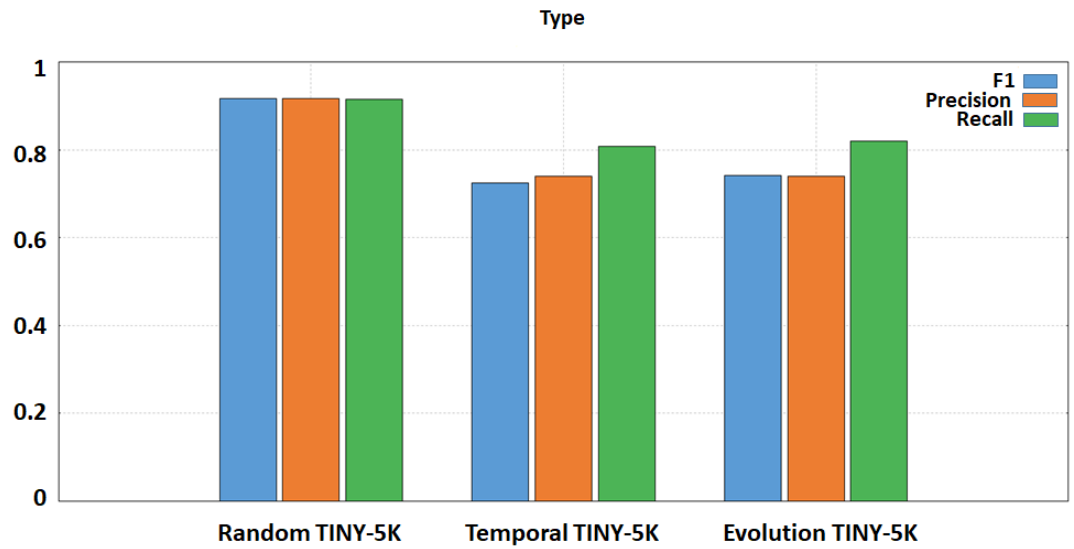
Comparison of SIR-GN Against ResNet18

- The ***Inferential SIR-GN*** approach came out on top by achieving a macro-F1 score of **0.718**, a macro-recall of **0.794**, and a macro-precision of **0.729** over **MALNET-TINY 61K** dataset.

Experimental Evaluation

SIR-GN Performance over MALNET-TINY 5K Dataset

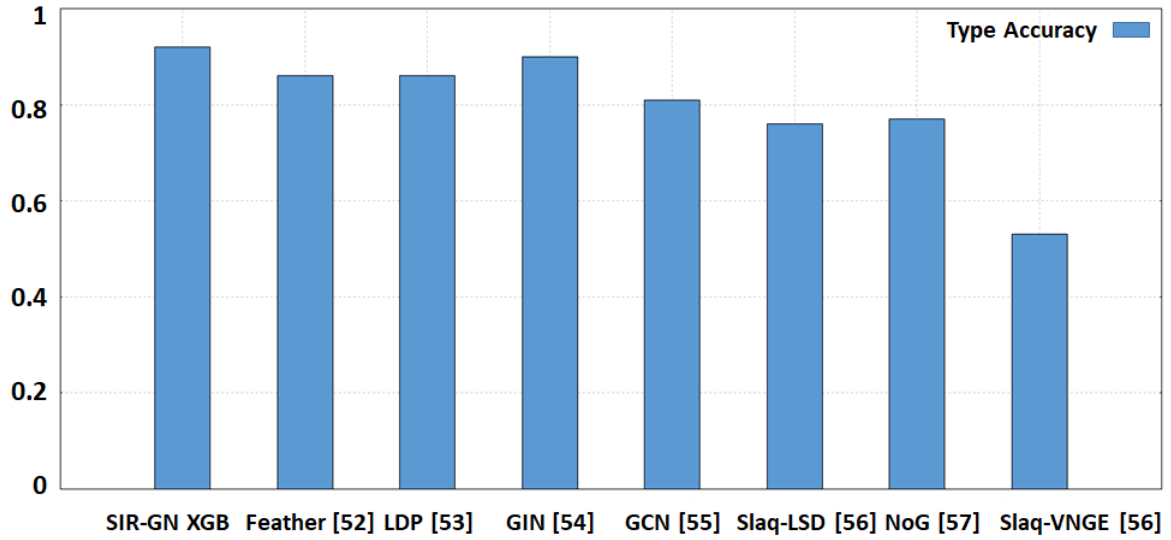




Experimental Evaluation

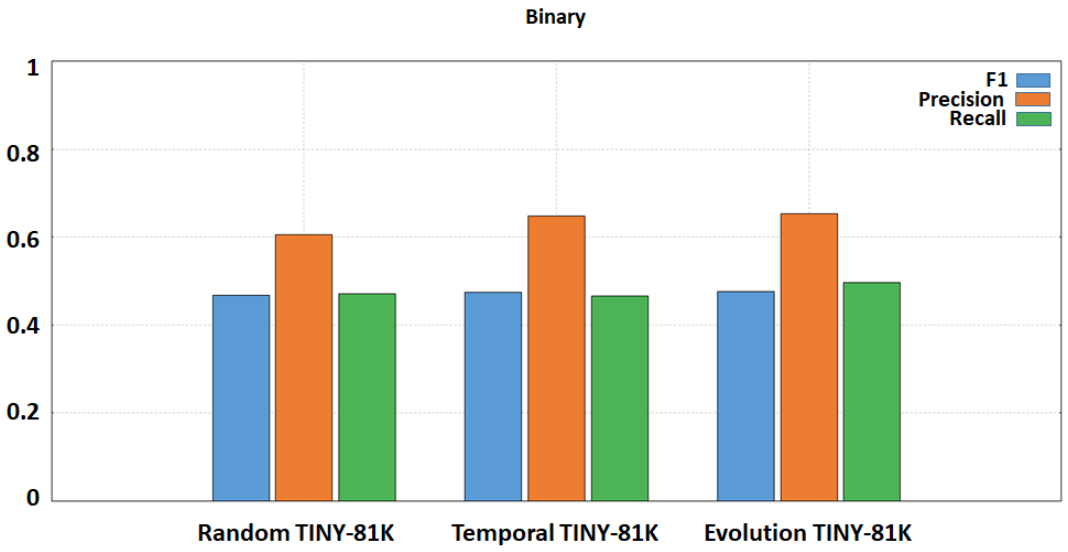
Comparison of SIR-GN Against Several Methods of the Literature

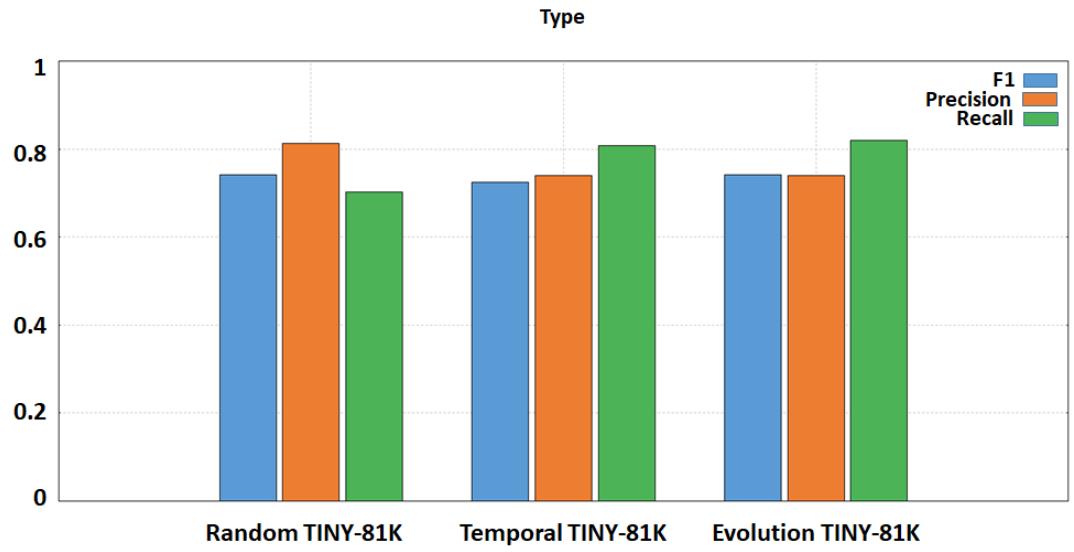
- The ***Inferential SIR-GN*** approach emerged as the highest-performing method, achieving an accuracy of 0.92 over **MALNET-TINY 5K** dataset.



Experimental Evaluation

SIR-GN Performance over MALNET-TINY 81K Dataset





Conclusion & Future Work

- The main contribution of this paper:
  - ✓ The proposition of an innovative approach for developing an Android malware classifier.
  - ✓ Training and testing the proposed model to recognize and classify malware as well as the polymorphic version of the malware.

- In future work and as a continuation of this research, we propose to extend our framework in order to handle with other complex Machine Learning approaches.

Thank You For Your Attention

SDS2023
IEEE SWISS CONFERENCE ON DATA SCIENCE

The 10th IEEE Swiss Conference on Data Science
June 22 – 23, 2023, Zürich, Switzerland

IEEE

A Graph-Representation-Learning Framework for Supporting Android Malware Identification and Polymorphic Evolution

Alfredo Cuzzocrea, Miguel Querbado, Abderraouf Hafsaoui, Edoardo Serra

iDEA Lab, University of Calabria, Rende, Italy

Computer Science Dept., Boise State University, Boise ID, USA