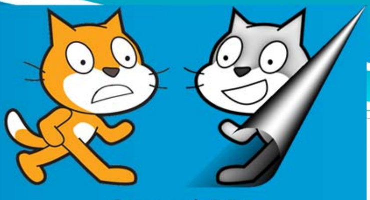
SCRATCH CODE CLONE DETECTION

Analyzing the Similarity of Scratch Programs

University of Alabama at Birmingham
Joshua Swank, Joel Tully, Brittany Stewart, Barrett Bryant
{joshuasw, kjoel, gtg826q, bryant}@cis.uab.edu

University of Alabama
Jeff Gray
gray@cs.ua.edu



ScratchPD
Catching the copy-cats

Software Composition and Modeling Laboratory
SoftCom
Department of Computer and Information Sciences
University of Alabama at Birmingham

THE UNIVERSITY OF ALABAMA
ENGINEERING

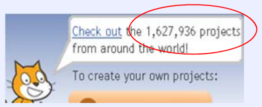
This material is based upon work supported by the National Science Foundation under Grant No. CCF-0702764

Introduction

Although Scratch is an excellent tool to teach computer science concepts, it is easy for students to plagiarize a Scratch assignment by downloading someone else's project from the internet. This poster describes our efforts in designing a tool (called the Scratch Plagiarism Detector - ScratchPD) that can compare Scratch programs and provide a similarity measure between a candidate program compared to a large set of existing programs. The poster motivates the need for such capability and offers a summary of the technical issues involved in performing program analysis on Scratch code.

Motivation

Scratch, as intended, has a large internet-based social community where users can share their projects with others. Since its development, the online Scratch community has grown large due to Scratch's success and popularity. As the community has grown, so has the collection of user-submitted programs.



Check out the 1,627,936 projects from around the world!
To create your own projects:

The availability of a large collection of existing projects presents a challenge for situations that expect the authenticity of student-submitted projects, such as: 1) graded assignments for a class project, and 2) various types of Scratch contests and festivals where originality and novelty are as important as creative extension. The challenge is that a student given an assignment can easily find and reuse a similar existing program among the vast collections online.

Is it allowed if someone takes your game, changes the theme, then calls it "their creation"? Because I created a game called <...> and a week later, a user called <...> redid the background, and called it "her creation" and I am really annoyed with her for taking credit for MY game.

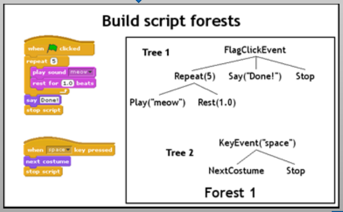
Approach

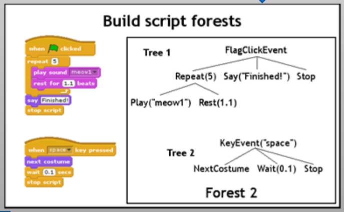
ScratchPD will quickly and effectively search a collection of Scratch programs and identify any projects that fall within a certain threshold of similarity with the project in question. Below is a summary of our approach for detecting similarity:

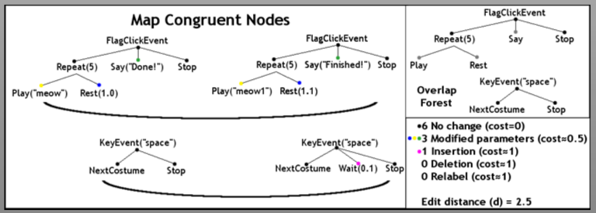
- Most of the focus thus far has been on the problem of comparing scripts. It was decided that the *edit distance* (i.e., the minimum number of changes needed to make one set of scripts identical to another) is a metric that describes the similarity of script sets in a meaningful way.
- In order to calculate the edit distance, a forest (set of trees) is constructed for each sprite. Every tree in the forest represents a single script. The most current algorithm for finding the edit distance of trees was put forward by Demaine, Mozes, Rossman, and Weimann and has a worst-case time complexity of $O(n^3)$. To compare entire forests, a variation of the forest comparison algorithm discussed by Zhang and Shasha is used.
- After the congruent nodes are mapped and the edit distance found, a percentage can be calculated to describe the similarity of any component of the project.
- A project is flagged if the similarity of any component over a certain size surpasses a threshold α. For example, the program could find a list of projects that have a script longer than 10 statements that is more than 50% similar to a script in the original project.

Example

Original submissions?







A percent uniqueness can now be found between the projects or any two congruent sprites or scripts. In this case, the largest project has 10 nodes.

$$\text{Similarity} = 1 - (\text{distance} / \text{max size}) = 0.75$$

So in this example, the projects are 75% similar.