

Usage and Attribution of

 **stackoverflow** Code Snippets

in **GitHub** Projects



Sebastian Baltes

 @s_baltes

 empirical-software.engineering

41st International Conference on
Software Engineering (ICSE 2019)
May 29-31, Montreal, Canada

Thanks to my co-author!

Empirical Software Engineering

<https://doi.org/10.1007/s10664-018-9650-5>



CrossMark

Usage and attribution of Stack Overflow code snippets in GitHub projects

Sebastian Baltes¹  · Stephan Diehl¹ 

Published online: 01 October 2018

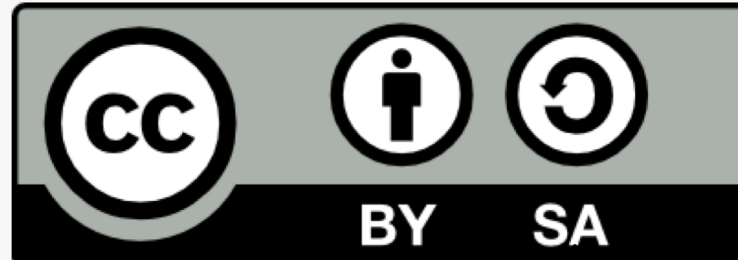
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

Stack Overflow (SO) is the most popular question-and-answer website for software developers, providing a large amount of copyable code snippets. Using those snippets raises maintenance and legal issues. SO's license (CC BY-SA 3.0) requires attribution, i.e., referencing the original question or answer, and requires derived work to adopt a compatible license. While there is a heated debate on SO's license model for code snippets and the

Stack Overflow's License

"You must give **appropriate credit** [...] and indicate if changes were made."



Attribution

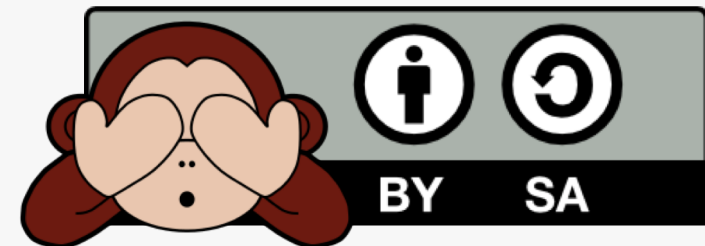
Share-alike

"If you [...] **build upon** the material, you must **distribute your contributions** under the same license as the original."

Results from our Online Surveys

- **46%** of the participants admitted copying code from Stack Overflow **without attribution**
- **75%** did **not know** that content on SO is licensed under **CC BY-SA**
- **67%** did **not know** that **attribution is required**

→ **Lack of awareness**



Background



“Well, but these snippets are rather trivial and not protected by copyright.”

- Not all code snippets on Stack Overflow are copyrightable
- “A snippet that is more than one or two lines of standard function calls would typically be creative enough for copyright” [Engelfriet 2016]
- But no “international standard for originality” [Creative Commons 2017b]



Here's what I do:

889

1. First of all I check what providers are enabled. Some may be disabled on the device, some may be disabled in application manifest.
2. If any provider is available I start location listeners and timeout timer. It's 20 seconds in my example, may not be enough for GPS so you can enlarge it.
3. If I get update from location listener I use the provided value. I stop listeners and timer.
4. If I don't get any updates and timer elapses I have to use last known values.
5. I grab last known values from available providers and choose the most recent of them.

Here's how I use my class:

```
LocationResult locationResult = new LocationResult(){
    @Override
    public void getLocation(Location location){
        //Got the location!
    }
};
MyLocation myLocation = new MyLocation();
myLocation.getLocation(this, locationResult);
```

And here's MyLocation class:

```
import java.util.Timer;
import java.util.TimerTask;
import android.content.Context;
import android.location.Location;
import android.location.LocationListener;
import android.location.LocationManager;
import android.os.Bundle;

public class MyLocation {
    Timer timer1;
    LocationManager lm;
    LocationResult locationResult;
    boolean gps_enabled=false;
    boolean network_enabled=false;

    public boolean getLocation(Context context, LocationResult result)
    {
        //I use LocationResult callback class to pass location value from MyLocat
        locationResult=result;
        if(lm==null)
            lm = (LocationManager) context.getSystemService(Context.LOCATION_SERV

        //exceptions will be thrown if provider is not permitted.
        try(gps_enabled=lm.isProviderEnabled(LocationManager.GPS_PROVIDER);}catch
        try(network_enabled=lm.isProviderEnabled(LocationManager.NETWORK_PROVIDER

        //don't start listeners if no provider is enabled
        if(!gps_enabled && !network_enabled)
            return false;

        if(gps_enabled)
            lm.requestLocationUpdates(LocationManager.GPS_PROVIDER, 0, 0, locatio
        if(network_enabled)
            lm.requestLocationUpdates(LocationManager.NETWORK_PROVIDER, 0, 0, loc
```

Somebody may also want to modify my logic. For example if you get update from Network provider don't stop listeners but continue waiting. GPS gives more accurate data so it's worth waiting for it. If timer elapses and you've got update from Network but not from GPS then you can use value provided from Network.

One more approach is to use LocationClient <http://developer.android.com/training/location/retrieve-current.html>. But it requires Google Play Services apk to be installed on user device.

share improve this answer edited Jun 25 '13 at 9:33 answered Jun 30 '10 at 0:07

Fedor 40k ● 9 ● 71 ● 86



```
public class MyLocation {
    Timer timer1;
    LocationManager lm;
    LocationResult locationResult;
    boolean gps_enabled=false;
    boolean network_enabled=false;

    public boolean getLocation(Context context, LocationResult result)
    {
        // Use LocationResult callback class to pass location value from MyLocation to user code.
        locationResult=result;
        if(lm==null)
            lm = (LocationManager) context.getSystemService(Context.LOCATION_SERVICE);

        //exceptions will be thrown if provider is not permitted.
        try(gps_enabled=lm.isProviderEnabled(LocationManager.GPS_PROVIDER);}catch(Exception ex){}
        try(network_enabled=lm.isProviderEnabled(LocationManager.NETWORK_PROVIDER);}catch(Exception ex){}

        //don't start listeners if no provider is enabled
        if(!gps_enabled && !network_enabled)
            return false;

        if(gps_enabled)
            lm.requestLocationUpdates(LocationManager.GPS_PROVIDER, 0, 0, locationListenerGps);
        if(network_enabled)
            lm.requestLocationUpdates(LocationManager.NETWORK_PROVIDER, 0, 0, locationListenerNetwork);
        timer1=new Timer();
        timer1.schedule(new GetLastLocation(), 20000);
        return true;
    }

    LocationListener locationListenerGps = new LocationListener() {
        public void onLocationChanged(Location location) {
            timer1.cancel();
            locationResult.getLocation(location);
            lm.removeUpdates(this);
            lm.removeUpdates(locationListenerNetwork);
        }
        public void onProviderDisabled(String provider) {}
        public void onProviderEnabled(String provider) {}
        public void onStatusChanged(String provider, int status, Bundle extras) {}
    };

    LocationListener locationListenerNetwork = new LocationListener() {
        public void onLocationChanged(Location location) {
            timer1.cancel();
            locationResult.getLocation(location);
            lm.removeUpdates(this);
            lm.removeUpdates(locationListenerGps);
        }
        public void onProviderDisabled(String provider) {}
        public void onProviderEnabled(String provider) {}
        public void onStatusChanged(String provider, int status, Bundle extras) {}
    };

    class GetLastLocation extends TimerTask {
        @Override
        public void run() {
            lm.removeUpdates(locationListenerGps);
            lm.removeUpdates(locationListenerNetwork);

            Location net_loc=null, gps_loc=null;
            if(gps_enabled)
                gps_loc=lm.getLastKnownLocation(LocationManager.GPS_PROVIDER);
            if(network_enabled)
                net_loc=lm.getLastKnownLocation(LocationManager.NETWORK_PROVIDER);

            //if there are both values use the latest one
            if(gps_loc!=null && net_loc!=null){
                if(gps_loc.getTime()>net_loc.getTime())
                    locationResult.getLocation(gps_loc);
                else
                    locationResult.getLocation(net_loc);
                return;
            }


            if(gps_loc!=null){
                locationResult.getLocation(gps_loc);
                return;
            }
            if(net_loc!=null){
                locationResult.getLocation(net_loc);
                return;
            }
            locationResult.getLocation(null);
        }
    }

    public static abstract class LocationResult{
        public abstract void getLocation(Location location);
    }
}
```





The screenshot shows three GitHub repository pages. The top one is 'WuhanMonkey/MoboSensAndroid' showing the file 'src/nano/Mobo/Sens/MyLocation.java'. The middle one is 'peruldem/DPR-KITA' showing 'app/src/main/java/id/gits/dprkita/utlis/MyLocation.java'. The bottom one is 'pacosal/ownmdm' showing 'src/com/pacosal/mdm/MyLocation.java'. Arrows from the Stack Overflow code snippet point to these files, indicating where the code was reused.

Stack Overflow Code in the OpenJDK

 JDK / JDK-8170860
Get rid of the humanReadableByteCount() method in openjdk/hotspot

Details

Type:	 Bug	Status:	RESOLVED
Priority:	 P2	Resolution:	Fixed
Affects Version/s:	9	Fix Version/s:	9
Component/s:	hotspot		

implement the method `humanReadableByteCount` which body was copied from the Stack Overflow site: <https://stackoverflow.com/a/3758880>

It's just a few lines of code, but it **could cause legal issues.** The method should be either re-implemented or removed.

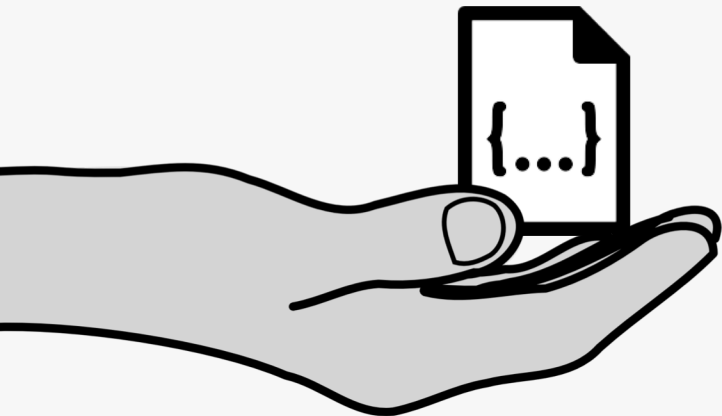
Besides the potential legal issues, duplicating a code is **not a good practice.**

<https://bugs.openjdk.java.net/browse/JDK-8170860>

Implications of Stack Overflow's License

Permissive Licenses

- Permit using the licensed source code in proprietary software **without publishing changes** or the **derived work**
- *Examples:* MIT, Apache, and BSD license families



Copyleft Licenses

- Requires either modifications to the licensed content or the complete derived work to be **published under the same or a compatible license** (share-alike)
- *Examples (weak copyleft):* Mozilla/Eclipse Public Licenses
- *Examples (viral copyleft):* GNU General Public Licenses, Creative Commons Share-Alike Licenses (e.g., **CC BY-SA**)

Enforceability of Copyleft Licenses

- Courts in the US and Europe ruled that open source licenses are **enforceable contracts**
- Authors are able to **sue** when terms such as the share-alike requirement are violated:
 - **Interdict distribution** of derived work
 - **Claim monetary damages**
- USA: DMCA takedown notices for allegedly infringed copyright
 - Example: <https://github.com/github/dmca>
- Risk in mergers and acquisitions of companies
 - Example: FSF vs. Cisco lawsuit



Research Question



Question:

How **frequently** is code from Stack Overflow posts used in public GitHub projects **without** the required **attribution**?

Approach:

Triangulate an estimate for the attribution ratio using three different methods.

Results

Rank	Matches				Recall	Attribution	
	ALL	DISTINCT	REF	NO-REF	REF/ F_{AQ}	REF/DISTINCT	F_{AQ} /DIST.
1	997	448	97	351	79.5%	21.7%	27.2%
2	1,843	913	60	853	60.0%	6.6%	11.0%
3	2,662	902	87	815	80.6%	9.6%	12.0%
4	420	170	18	152	94.7%	10.6%	11.2%
5	1,492	402	25	377	73.5%	6.2%	8.5%
6	2,642	807	65	742	87.8%	8.1%	9.2%
7	160	124	12	112	29.3%	9.7%	33.1%
8	355	174	22	152	61.1%	12.6%	20.7%
9	295	225	5	220	10.6%	2.2%	20.9%
10	65	33	11	22	42.3%	33.3%	78.8%
All	10,931	4,198	402	3,796	<i>M</i> 61.9%	<i>M</i> 12.1%	<i>M</i> 23.2%

Method 2: Code Clone Detector

- **Goal:** Use code clone detector to find clones of a sample of Stack Overflow snippets in a sample of GitHub projects

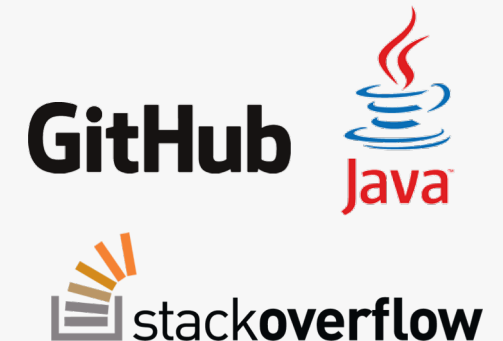
- *Why samples?*

- Code clone detection is computationally expensive



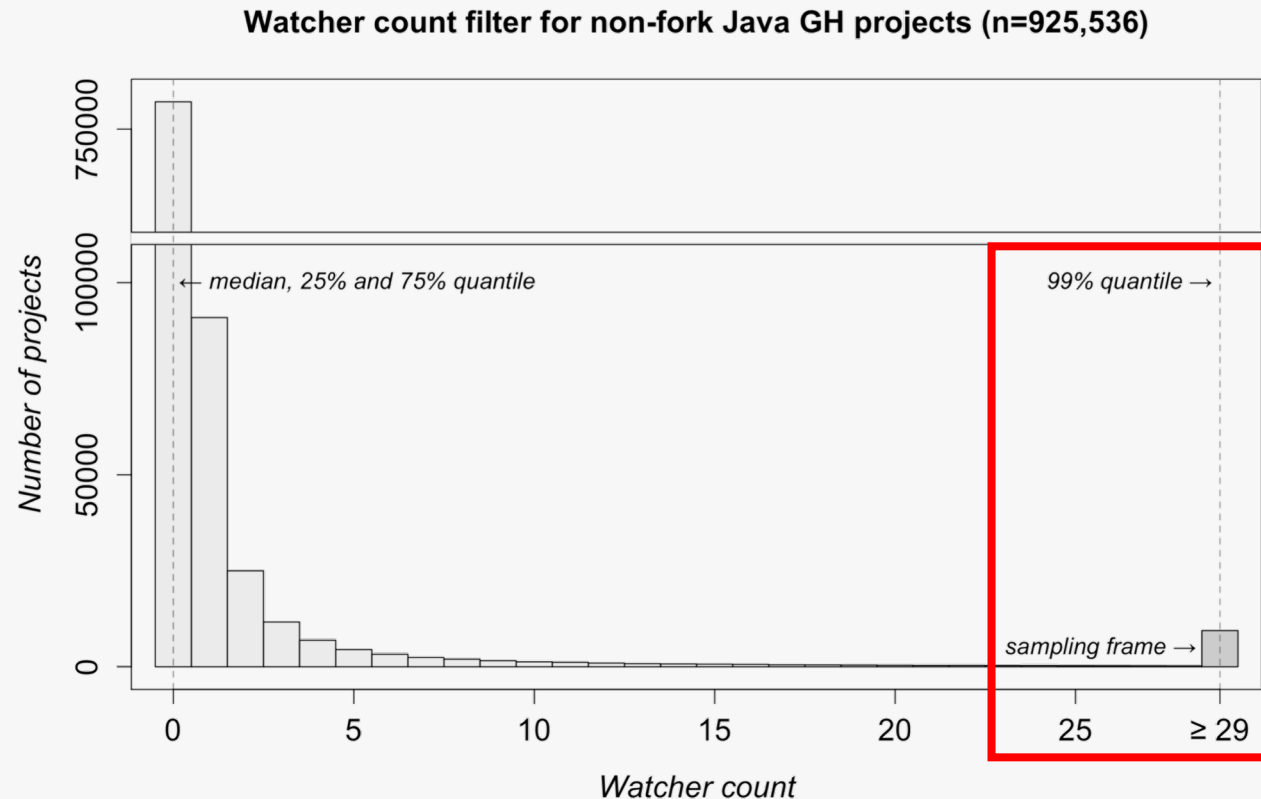
- *Which snippets and projects to select?*

- Random samples: Many **toy projects** on GitHub and many **irrelevant snippets** on Stack Overflow
- Purposive sampling: Limited generalizability



GitHub Project Sample

- Focus on **popular** GitHub projects
- High precision in selecting “engineered” software projects [Munaiah et al. 2017]
- Greater (potential) impact of licensing issues



Sample size:
3,000 / 2,313



Stack Overflow Snippet Samples

- Non-trivial snippets retrieved from 100 most frequently referenced answers (n=111)
⇒ S_{top100}
- Non-trivial snippets retrieved from answers referenced in GitHub projects (n=137)
⇒ S_{gh}
- *External sources:* Only three snippets available under a more permissive license than CC BY-SA

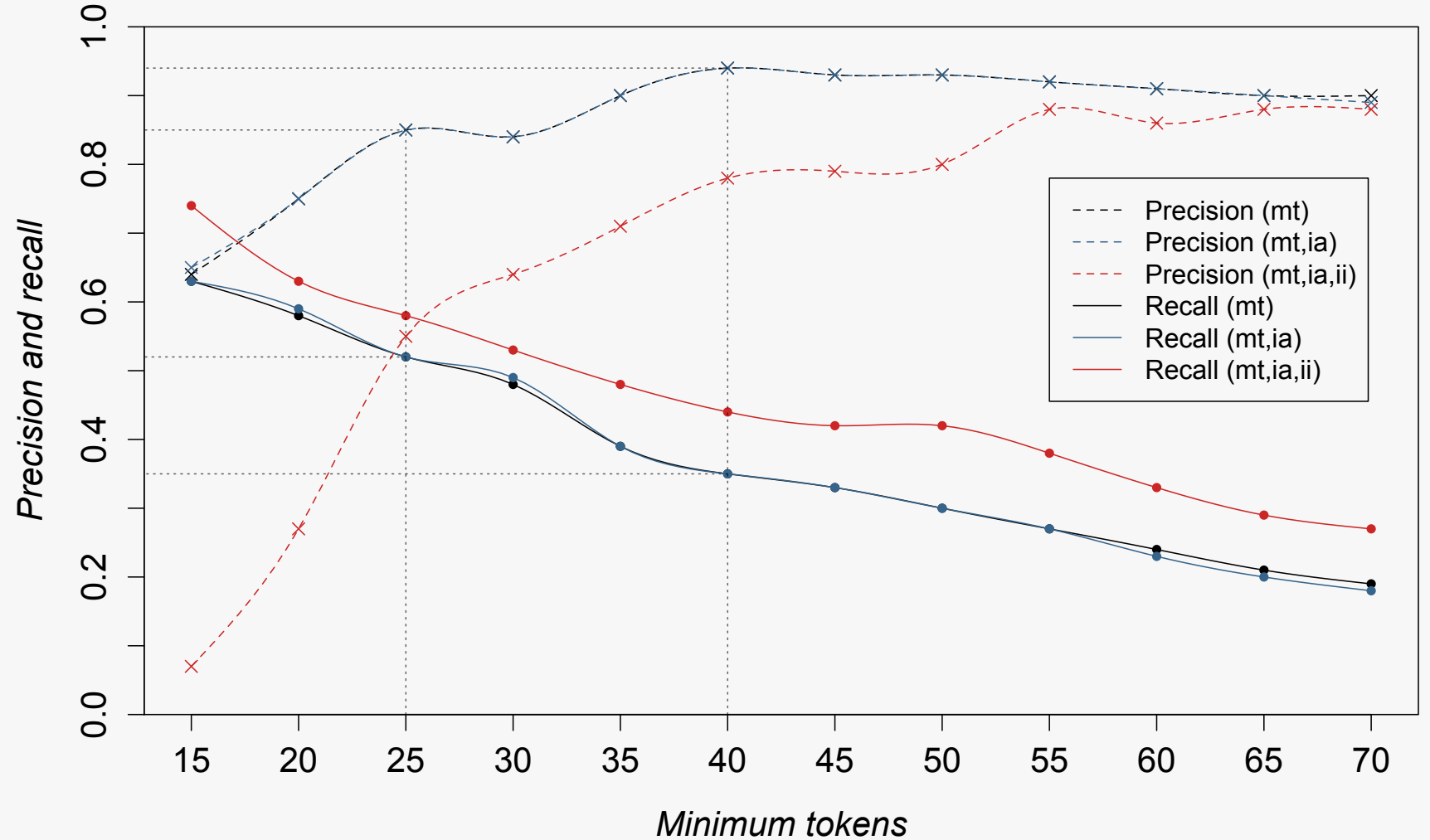


Code Clone Detector Calibration



<https://pmd.github.io/>

Comparison of CPD configurations



Results

Set	Snippets				Files		Repos
	ALL	MATCHED	ANSWERS	MATCHED	MATCH.	REF	MATCHED
S_{gh}	137	53 (39%)	102	52 (51%)	163	58 (36%)	124 (5%)
S_{top100}	111	48 (43%)	85	46 (54%)	173	25 (14%)	125 (5%)
US	222	101 (46%)	169	86 (51%)	297	70 (24%)	199 (9%)

Method 3: Exact Matches

- **Goal:** Address shortcomings of Method 1 and 2
 - Increase sample sizes
 - Exclude snippets available on external sources
 - Systematically exclude short snippets
- Select as many projects and snippets as possible and search for (almost) exact matches



Method 3: Exact Matches


 Google Cloud

GitHub 

209m files in
4.1m projects

- ✓ Project is not a fork, has ≥ 5 Java files and ≥ 1 watcher(s)
- ✓ File has ending `.java` has ≥ 68 NLOC (Q_3)




1.7m Java files
in 64k projects

Normalization and
substring search



10,358
matches

 Google Cloud

 **stackoverflow**

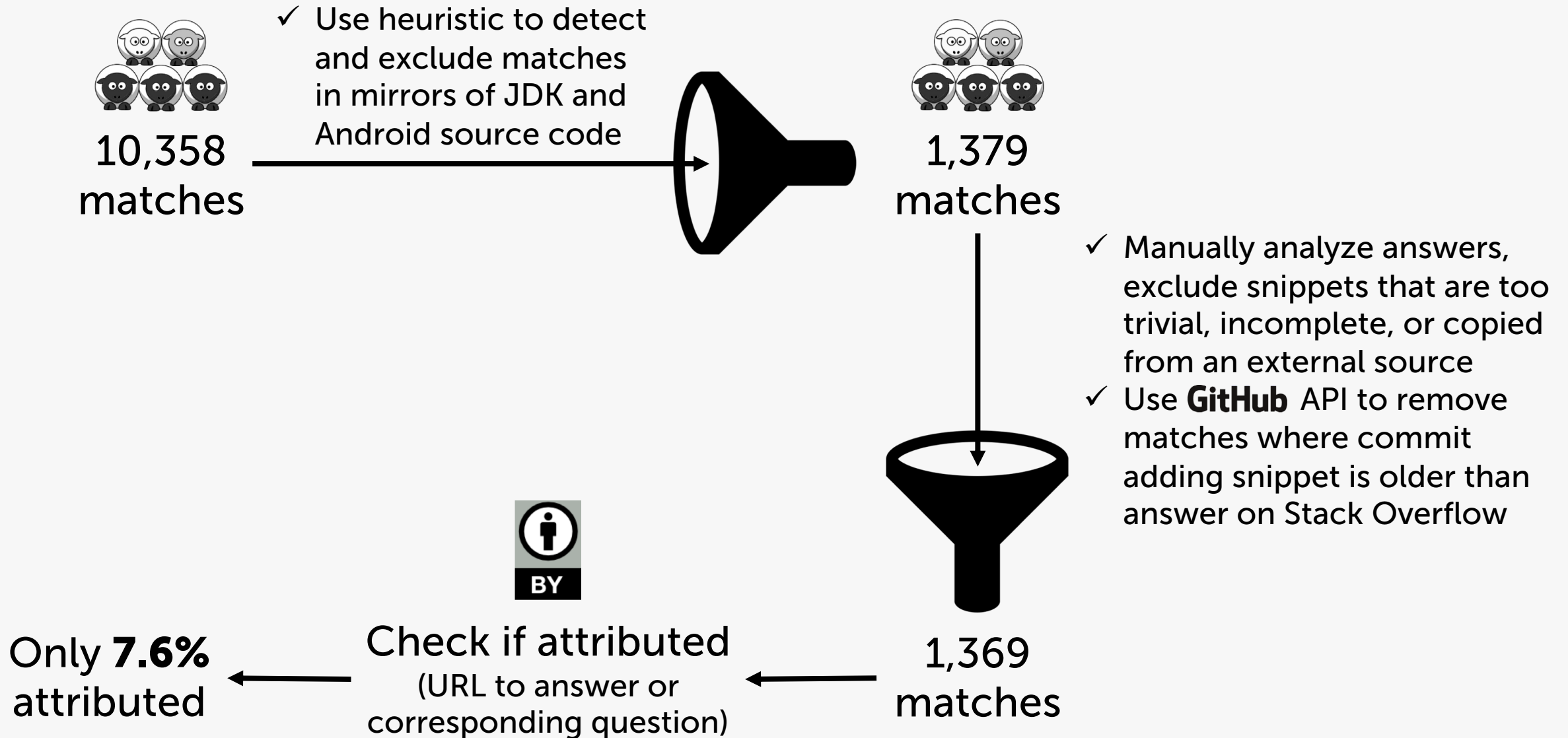
21m answers

- ✓ Question tagged `java` or `android`
- ✓ Answer score ≥ 10
- ✓ Code block ≥ 6 NLOC




29k snippets from
24k answers

Method 3: Filtering of Matches



Attribution



Attribution ratio:

- Method 1 (regular expressions): 23 %
- Method 2 (code clone detector): 24 %
- Method 3 (exact matches): 8 %

Conservative estimate:

- **Attribution ratio \leq 25%**

Share-alike



Only **2%** of all analyzed repositories (all methods) containing code from Stack Overflow **attributed** its source and used a **compatible license** (not CC BY-SA, but GPL 3.0).

SPDX license name	Number of repos containing a SO code snippet clone that was:	
	unattributed (<i>n</i> = 2,962)	attributed (<i>n</i> = 329)
Apache-2.0	921 (31.1%)	99 (30.1%)
MIT	621 (21.0%)	72 (21.9%)
GPL-3.0	435 (14.7%)	60 (18.2%)
GPL-2.0	284 (9.6%)	21 (6.4%)
BSD-3-Clause	82 (2.8%)	9 (2.7%)

Method 1

SPDX license name	Number of repos containing a SO code snippet clone that was:	
	unattributed (<i>n</i> = 144)	attributed (<i>n</i> = 55)
None	56 (38.9%)	18 (32.7%)
Apache-2.0	33 (22.9%)	15 (27.3%)
GPL-3.0	17 (11.8%)	6 (10.9%)
MIT	6 (4.2%)	4 (7.3%)
GPL-2.0	4 (2.8%)	2 (3.6%)

Method 2

SPDX license name	Number of repos containing a SO code snippet clone that was:	
	unattributed (<i>n</i> = 1,169)	attributed (<i>n</i> = 163)
Apache-2.0	353 (30.2%)	36 (37.4%)
MIT	239 (20.4%)	25 (15.3%)
GPL-3.0	211 (18.0%)	19 (11.7%)
None	153 (13.1%)	61 (37.4%)
GPL-2.0	89 (7.61%)	8 (4.9%)

Method 3

Reaching out to Developers

- **Contacted owners** of GitHub projects containing copies of Stack Overflow snippets
- **75% not aware** of CC BY-SA licensing
(see slide about online surveys)
- Many thankful responses



Usage and Attribution of

 **stackoverflow** Code Snippets



in **GitHub** Projects

Sebastian Baltes
 @s_baltes

snippets.sbaltes.com
Data and scripts available on Zenodo

