An Empirical Study of License Violations in Open Source Projects

Arunesh Mathur[¶] **Harshal Choudhary**[¶] **Priyank Vashist**[¶]

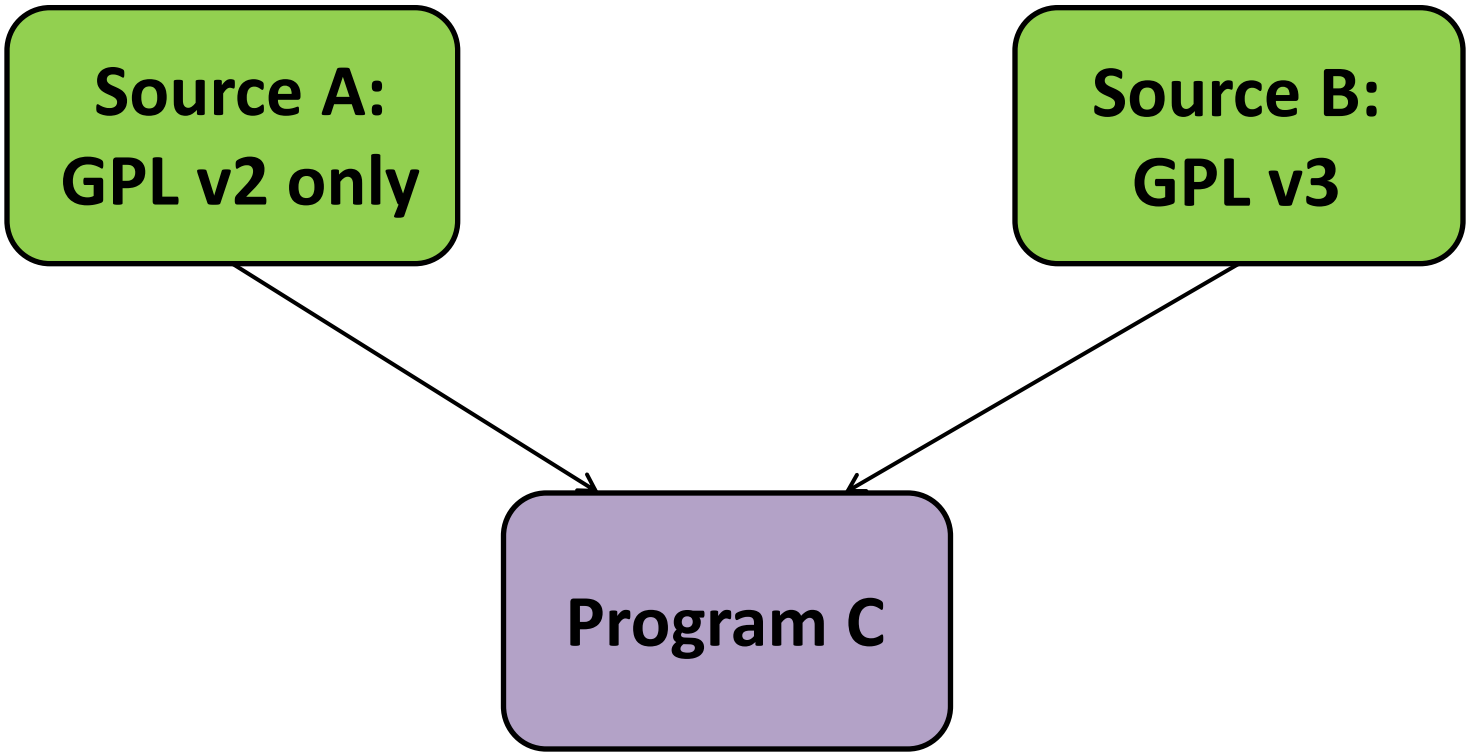
William Thies[†] **Santhi Thilagam**[¶]

[†] Microsoft Research India

[¶] National Institute of Technology Karnataka, Surathkal

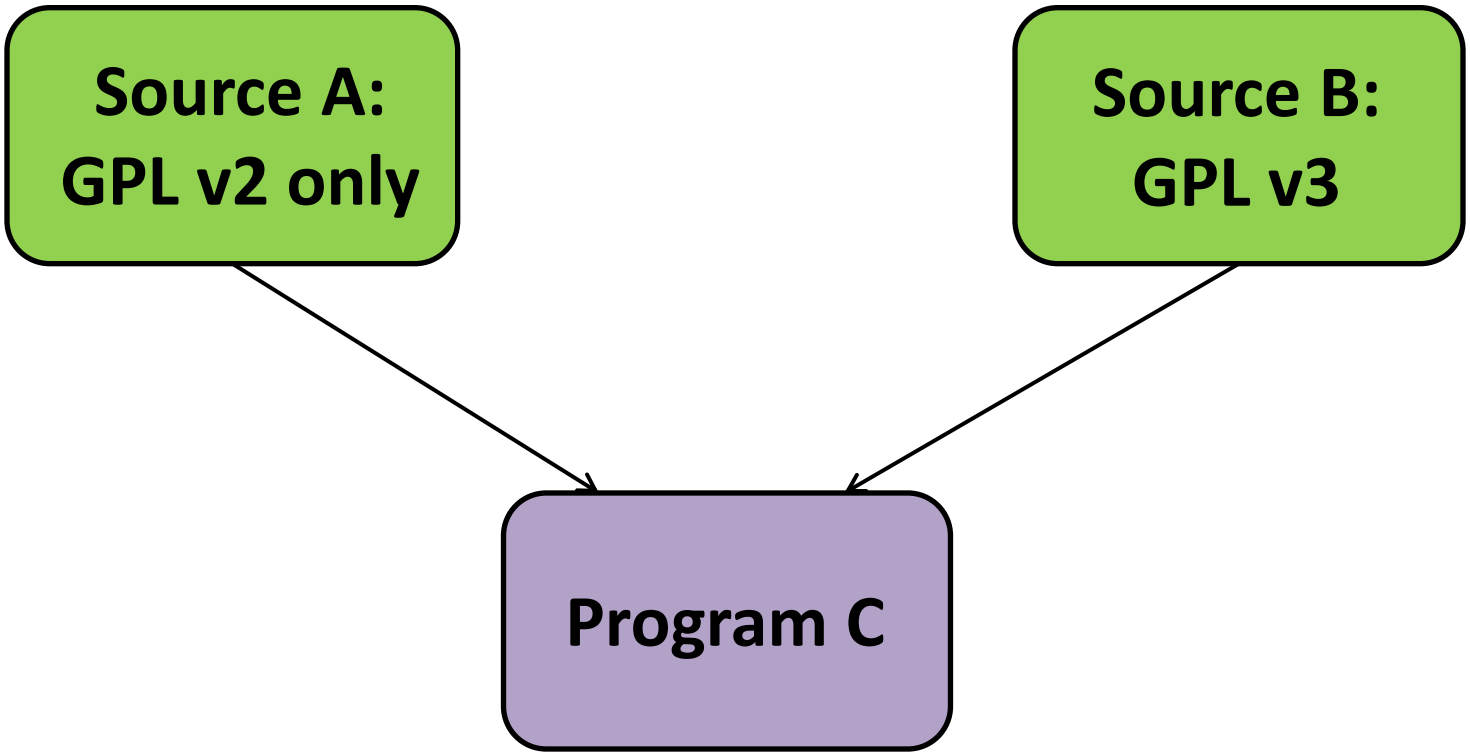
Source: http://www.groklaw.net/article.php?story=20100803132055210

Question



Source: http://www.groklaw.net/article.php?story=20100803132055210

Question



Is this valid?

Source: http://www.groklaw.net/article.php?story=20100803132055210

The BusyBox GPL violation (1/2)

- **GPL v2 licensed minimal Unix-like shell utilities optimized for use in embedded devices**
- **Have filed multiple cases of unlawful use; most recently against the likes of:**
  - Best Buy, Samsung, Westinghouse
  - JVC, Western Digital, Robert Bosch
  - Phoebe Micro, Humax USA
  - Comtrend, Dobbs-Stanford
  - Versa Technology, Zyxel Communications
  - Astak, GCI Technologies

Source: http://www.groklaw.net/article.php?story=20100803132055210

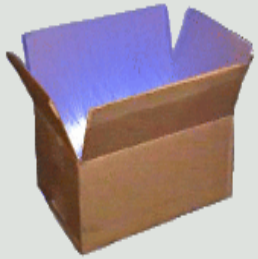
The BusyBox GPL violation (2/2)

- **What went wrong?**
  - Violated the GPL v2 by distributing the BusyBox binary as part of their products without the source code

- **Implications for one of the offenders:**
  - Damages worth $90,000
  - Lawyers' costs and fees worth $47,865
  - Donate all their infringing products in possession to charity

Source: http://www.busybox.net/shame.html

BUSYBOX



About

- About BusyBox
- BusyBox in VM
- Screenshot
- Announcements

Documentation

- FAQ
- Command Help

Get BusyBox

- Download Source
- License
- Products

Development

- Browse Source
- Source Control
- Mailing Lists
- Bug Tracking
- Contributing

Links

Hall of Shame!!!

This page is no longer updated, these days we forward this sort of thing to the Software Freedom Law Center instead.

The following products and/or projects appear to use BusyBox, but do not appear to release source code as required by the BusyBox license. This is a violation of the law! The distributors of these products are invited to contact Erik Andersen if they have any confusion as to what is needed to bring their products into compliance, or if they have already brought their product into compliance and wish to be removed from the Hall of Shame.

Here are the details of exactly how to comply with the BusyBox license, so there should be no question as to exactly what is expected. Complying with the Busybox license is easy and completely free, so the companies listed below should be ashamed of themselves. Furthermore, each product listed here is subject to being legally ordered to cease and desist distribution for violation of copyright law, and the distributor of each product is subject to being sued for statutory copyright infringement damages of up to $150,000 per work plus legal fees. Nobody wants to be sued, and Erik certainly would prefer to spend his time doing better things than sue people. But he will sue if forced to do so to maintain compliance.

Do everyone a favor and don't break the law -- if you use busybox, comply with the busybox license by releasing the source code with your product.

- Tritton Technologies NAS120
  see here for details
- Macsense HomePod
  with details here
- Compex Wireless Products
  appears to be running v0.60.5 with Linux version 2.4.20-uc0 on ColdFire, but no source code is mentioned or offered.
- Inventel DW 200 wireless/ADSL router
- Sweex DSL router
  appears to be running BusyBox v1.00-pre2 and udhcpd, but no source code is mentioned or offered.
- TRENDnet TEW-410APB

Software Licenses

- **Purpose:**
  - Means of using/distributing/modifying software without violating copyright laws
  - Protect the original author's rights
  - Have an effect on the end user's rights
- **Two types:**
  - Proprietary licenses
  - Free and Open Source (FOSS) licenses

Open Source Software (OSS) Licensing

- **Total of 69 Open Source Initiative (OSI) approved licenses (as of September 2012)**
  - Every open source license must follow the requirements listed in the Open Source Definition (OSD)
- **Varying flexibility of each license**
  - Has an impact on the degree of code reuse
  - Problems arise when merging components with incompatible licenses

Understanding Copyleft

- **Copyright** is the law by which an individual posses all rights to modify, distribute or copy his/her work
- **Copyleft** is the transfer of Copyright under the condition that the same rights are preserved in all future distributions/modifications (share-alike)

OSS License types

- **Three types:**
  - Strong Copyleft licenses
  - Weak Copyleft licenses
  - Permissive licenses
- **Copyleft licenses are “viral” in nature**
  - require the licensee to distribute the modified or derived work under the same license
  - Minimize the freedom to create software proprietary in nature

Open Source Software (OSS) Licensing

Strong Copyleft	Weak Copyleft	Permissive
GPL Free Software Free as in Freedom	LGPL Free Software Free as in Freedom	Apache
	Mozilla Public License	BSD
	eclipse	MIT

Goal of this Study

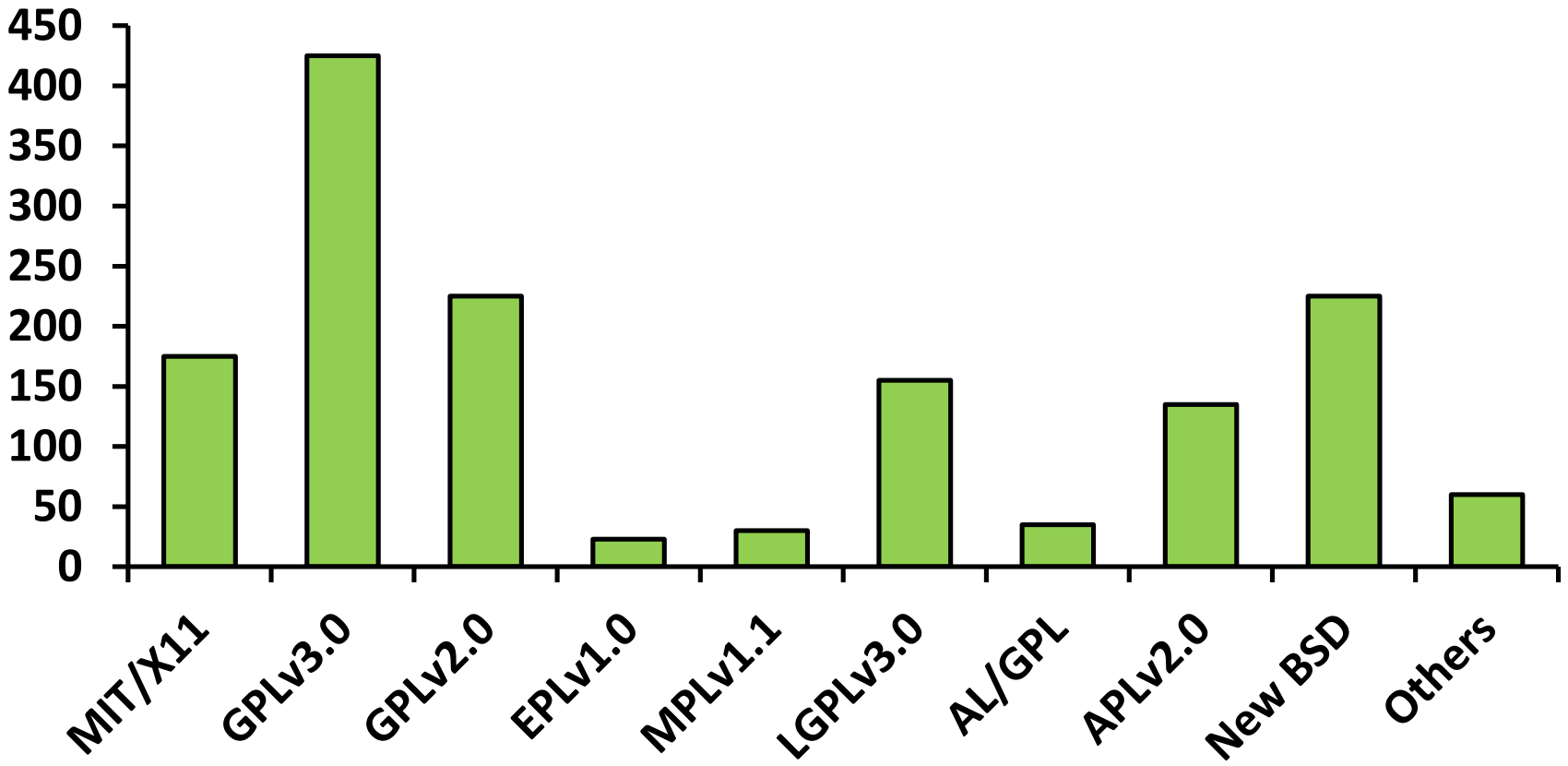
Colloquial evidence suggest that open source developers have a hard time with licenses as well

Aim to discover cases of violations in a large corpus of open source projects

Sample Set Selection

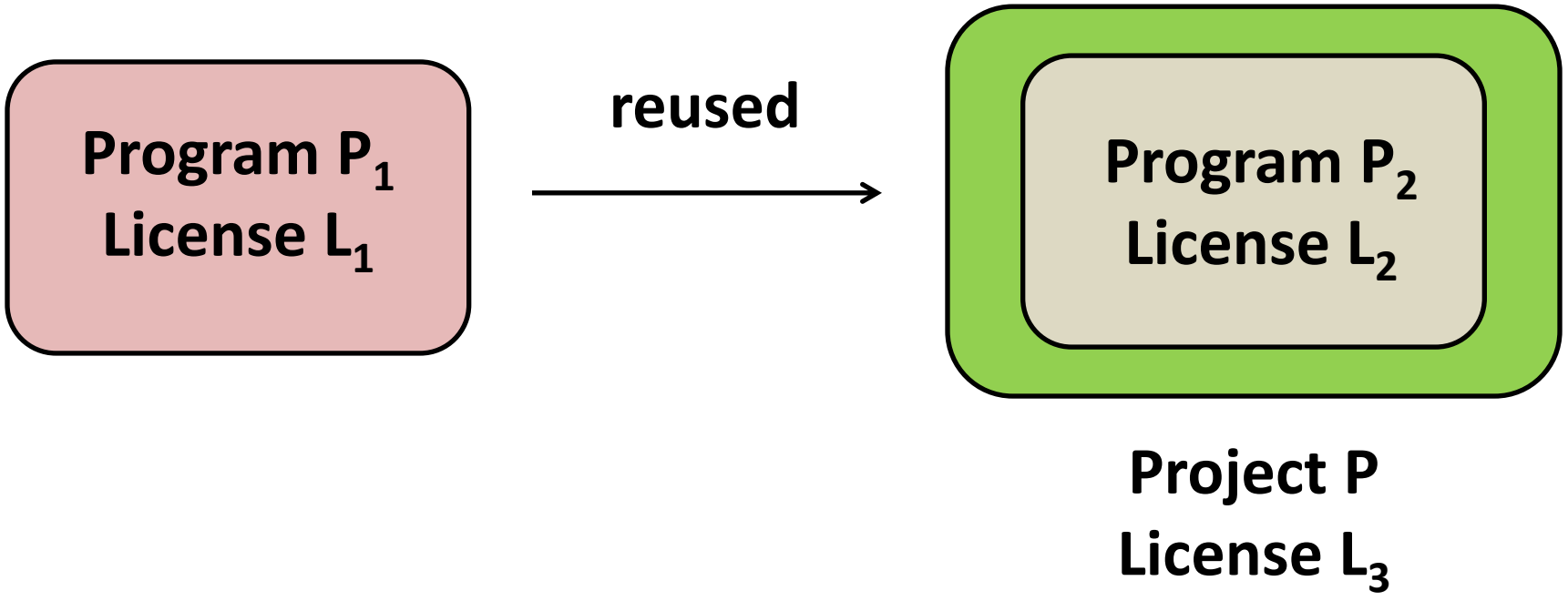
- **Retrieved a sample set of open source projects for examination**
  - 1423 open source projects from Google Code project hosting (http://code.google.com/hosting)
- **Random selection of sample space**
  - To get a good mix of project types, selected projects based on tags such as – *C, C++, Python, Java, Web, Flash, Embedded, Graphics, Android* etc.

Sample Set License Types



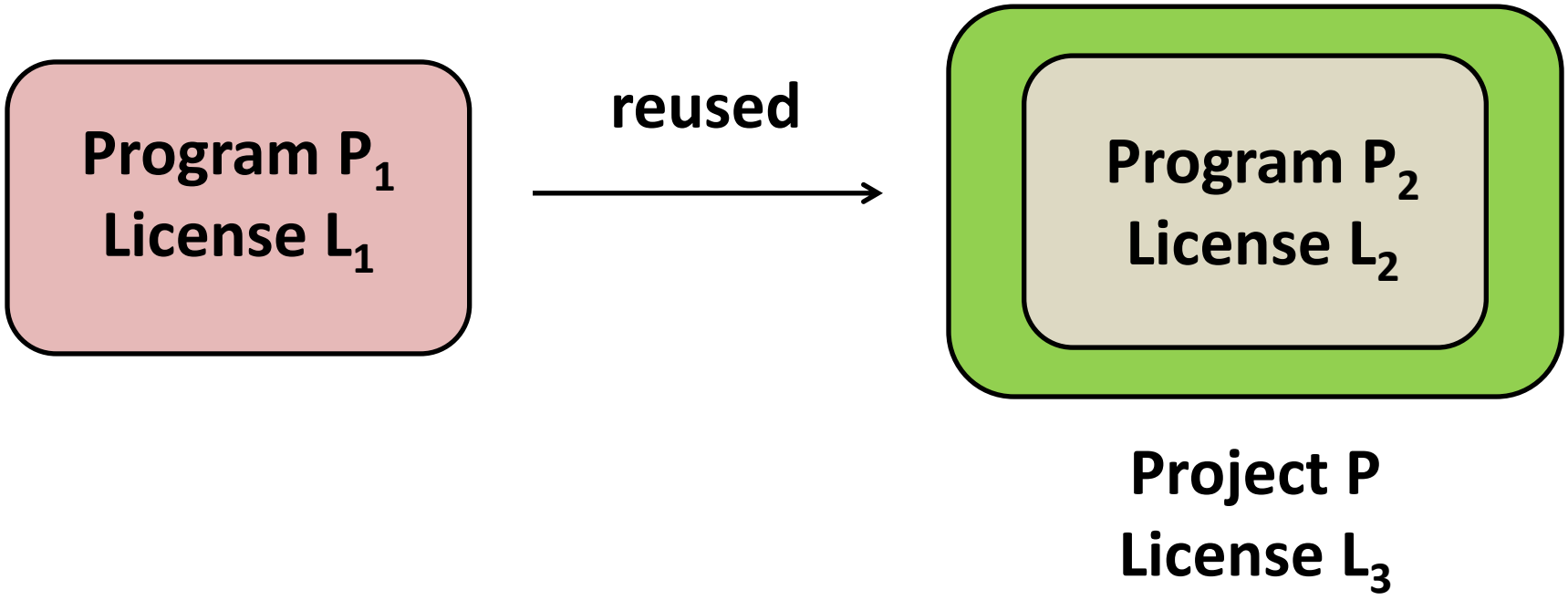
GPL v3.0 and GPL v2.0 ~ 40%

Defining Violations



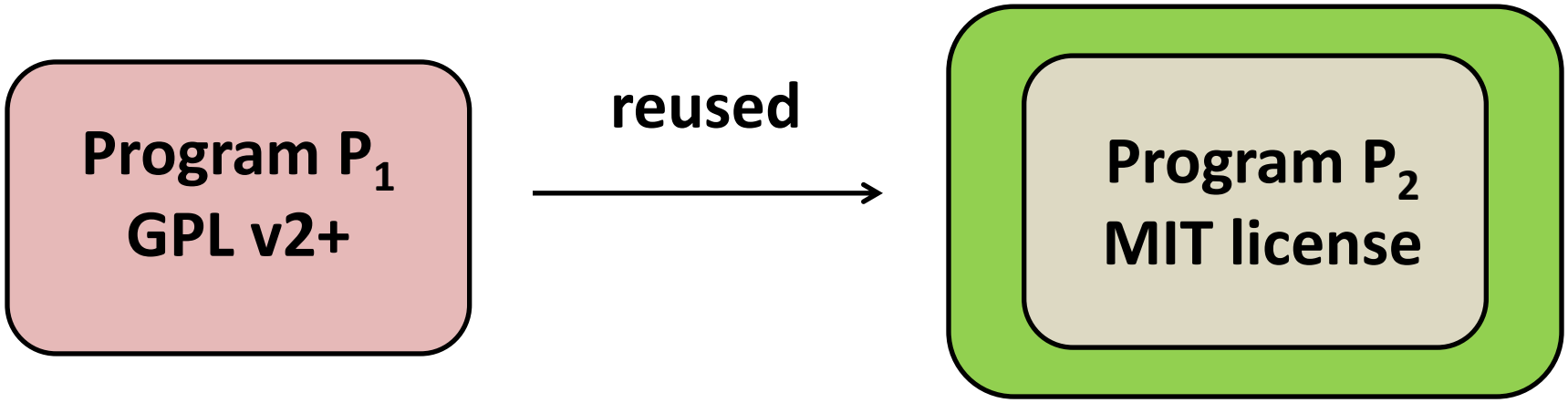
P_2 includes P_1 and derived works, if any

Defining Violations



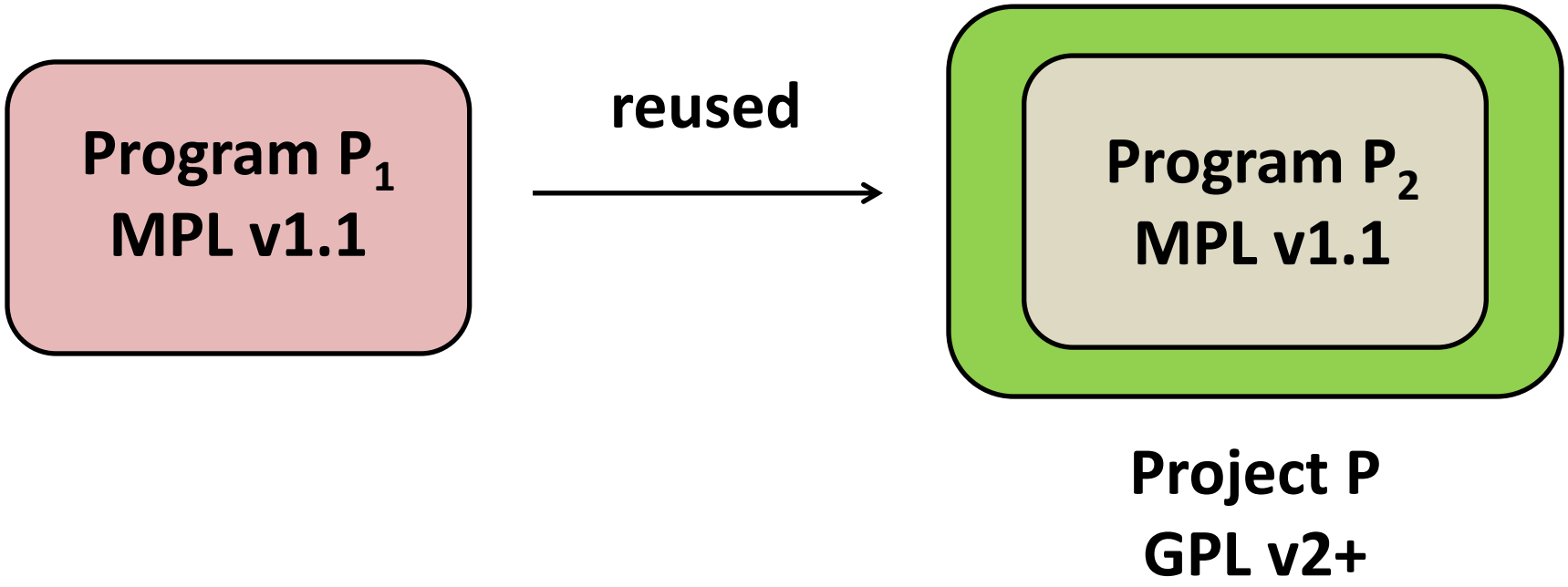
1. Check compatibility between L_1 and L_2
2. Check compatibility between L_2 and L_3

Defining Violations



GPLv2+ requires all derived/modified work (P_2) to be released under the same license

Defining Violations



GPLv2+ and the MPLv1.1 are incompatible

Detecting Code Reuse (1/3)

- To discover instances of code reuse, we use the ideas behind MOSS [Measure of Software Similarity], a plagiarism detection tool

- **Three step process:**
  - Preprocessing
  - Fingerprinting
  - Comparing

Detecting Code Reuse (2/3)

- **Preprocessing** phase removes unnecessary noise and unwanted characters in the source files
- **Fingerprinting** phase generates hashes after diving the preprocessed files into k-grams (strings of size *k*)
  - Size of *k* is programming language dependent
  - Hashing must minimize collisions

Detecting Code Reuse (3/3)

- **Comparison** phase groups files that have similar hashes together
  - #(hashes) for two files to be considered similar dependent on a threshold value
- To reduce false positives, we ignore hashes that correspond to license headers
- Pretty print files that are reported to be similar and manually examine them

Results (1/2)

- **Code Reuse:**
  - Discover a total of 103 cases of code reuse
  - Projects that have *High* activity are reused more than projects with *Medium* and *Low* activity
- **License Violations:**
  - 4 cases of license violations
  - GPL v2 being violated 3/4 times

Results (2/2)

Provider	Provider License	Acceptor	Acceptor License	Fix	Downloads
Miranda	GPL v2+	*TopToolBar*	LGPL v3+	Convey under GPLv3+	126
Miranda	GPL v2+	*Wi2Geoplugin*	MIT	Convey under GPLv2+	91,146
FLV Player	MPL v1.1	*Khan Academy*	Other Open Source	Choose compatible license	—
Arduino	GPL v2+	*Micropendous*	MIT	Keep parts under same license	1,238

Impact

- Exchanged emails with the developers of the violating projects
- *Micropendous* has since then, changed its license to GPL v2+ & MIT
- Developers of *Khan Academy* have acknowledged the lack of a license on their GitHub account
- Awaiting response from the rest

Conclusions

- **License compatibility turning into an intricate scenario**
  - Legal implications may have far reaching consequences for both – OSS and proprietary software developers
- **Multi-licensing**
  - Release under multiple licenses, if possible, to offer a wider choice to end users
- **Avoid forming new licenses to avoid dealing with existing ones upfront**

Acknowledgements

- Tom Callaway, Gervase Markham, Clint Adams and members of the apache-legal mailing list for useful discussions on open source licenses

- Supported by Microsoft Research India Travel Grant