Armijn Hemel
Tjaldur Software Governance Solutions

February 13, 2012

# About Armijn

- using Open Source software since 1994
- MSc Computer Science from Utrecht University (The Netherlands)
- core team `gpl-violations.org` since 2005
- ex-board member at NLUUG (`http://www.nluug.nl/`)
- owner Tjaldur Software Governance Solutions

# Subjects

- very brief overview of license violations
- problems in binary code clone detection
- open questions in binary code clone detection

# License enforcement

- Europe (Germany, France) & USA
- focus is on GPLv2 and LGPLv2/2.1
- done by companies (Nokia, Red Hat) and individual developers and projects (Harald Welte, BusyBox, XviD, etc.)

It is about copyright, not about patents!

# gpl-violations.org

Founded in 2004 by Harald Welte (copyright holder in the Linux kernel) to take on GPL license violations by:

- education
- documentation
- legal action

I have been active with `gpl-violations.org` since 2005.

So far we've had several hundred cases (most of them settled) and fixed many more using informal pressure.

# How gpl-violations.org works

1. we get a report via private email, public mailing list, chat, rumours, SMS, or our own research
2. if there is reasonable doubt about compliance of a device we do a test purchase to confirm the violation
3. if we confirm a violation we send a "cease and desist"

There are many false reports: a lot of people don't understand the license(s).

Our main focus is on consumer electronics (one of the biggest markets out there).

# Consumer electronics: the truth

Almost everything is purchased. Making everything yourself is commercial suicide:

- extremely thin margins
- cut throat competition
- quality is less important than price
- (ultra) short term thinking: companies don't know if they will still be in business in 6 months from now
- "cowboys"

It's like Nike: don't do any production, just marketing and sales.

In my experience typically more than 95% (or more) is reuse of open source software (with/without modifications)


See larger image: 300M 11N WIFI Router

Add to My Favorites

# 300M 11N WIFI Router

| | |
|---|---|
| FOB Price: | US $10 - 12 / Unit<br>Get Latest Price |
| Port: | Yantian |
| Minimum Order Quantity: | 100 Unit/Units |
| Supply Ability: | 50000 Piece/Pieces per Month |
| Payment Terms: | L/C,D/A,D/P,T/T,Western Union |
| Sample or Mini-Order: | Order now via ESCROW Buyer Protection |

Ms. Wiley Tsai
Offline

Contact Supplier
Send a Message to this Supplier

## Supplier Details

**Shenzhen Century Xinyang Tech Co., Ltd.**
[ Guangdong, China (Mainland) ]
Business Type:
Manufacturer
Contact Details
Gold Supplier [3rd Year]
A&V Checked

Online Showroom: 1,981 Products
510 Similar Products from this Supplier
View this Supplier's Website

Report Suspicious Activity

Product Details | Company Profile

## Quick Details

| | | |
|---|---|---|
| **Products Status:** Stock | **Type:** Wireless | **Application:** Soho |
| **Function:** Firewall, VPN | **LAN Ports:** 4 | **WAN Ports:** 1 |
| **Certification:** FCC, ROHS | **Brand Name:** Tianhao wifi router | **Model Number:** TH-R300M2 wifi router |
| **Place of Origin:** Guangdong China (Mainland) | **VPN:** Yes | **Number Of Ports:** 4 |
| **Antenna:** 2dBi with SMA port | **Chipset:** Ralink 3052 | **Function:** Supports DDWRT or OPEN DDWRT |

## Packaging & Delivery

**Packaging Detail:** Neutral color box 1pc/ color box 20pcs/ carton

# Problem source: supply chain

License violations are often a direct result of a mistake made in the supply chain:

- chipset vendors
- board makers
- SDK ("Software Development Kit") vendor
- reference design makers
- product customizers
- "labellers"

The "labellers" get sued and are responsible, even though they add/modify the least amount of code!

# Industry responses to enforcement

- extreme levels of frustration (problem doesn't go away by throwing money at it)
- they don't care about licenses, they just want to sell a product. Licenses are a nuisance that needs to be dealt with.
- a single enforcement case will make no change to the market (it is too big: a single company getting in trouble is not significant to push for change)
- no ill will. Companies want to fix it and there is a need for tools (cheap, or free) to do "due dilligence"

# Tools

Apart from the obvious “industry standard ”tools that solve some problems Tjaldur Software Governance Solutions has worked on tools to help solving specific problems in this field.

Goal: let companies do checks themselves, increasing quality and lowering costs.

- Binary Analysis Tool (Apache 2 license, freemium model)
- license scanning tools (leveraging existing tools like Ninka and FOSSology)
- long term: build system integration (preliminary work has been done)

# Binary Analysis Tool

- generic extensible pluggable framework for analysing binaries
- binary code clone detection using string comparisons: first extract string constants from the binary, compare it with a large database of data extracted from source code, finally assign a score to packages based on matches

Demo later this week.

# Academic research

There is a lot of unclarity about licenses and software provenance, and academia can help here. I'm trying to do my share, team up with researchers for research:

- "Finding Software License Violations Through Binary Code Clone Detection" (Mining Software Repositories 2011) - some results have been integrated into Binary Analysis Tool
- "What Goes into an Executable? Identifying a Binary's Sources by Tracing Build Processes" (sent to WCRE 2011 and ICSE 2012, unfortunatey rejected) - I'm preparing tooling that implements this
- future research (your name here)

# Open questions/problems

- detecting obfuscated code in binaries (when basic string comparisons simply aren't enough)
- detecting language embedding (interpreters, DSLs) in binaries (if they have been compiled)
- correlating binary code and source code (solved for source to binary using tracing, not from binary to source side)
- complete provenance of binary and source code files, down to the level of single commits (example: individual Git commits) because snapshots from DVCS (like Git) are rapidly replacing normal releases
- reducing false positives in detection: false claims can lead to counter lawsuits, with significant risks