“This needs to be fixed”

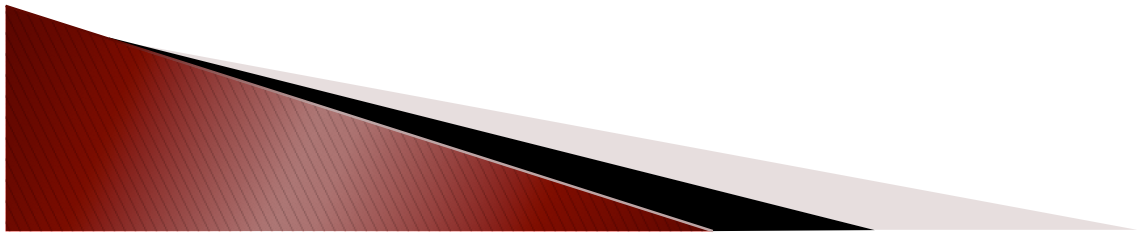
Jokes, vulnerabilities, and analysis of commit statements

Logan Lodge and Bruce Potter

l0l0@shmoo.com gdead@shmoo.com

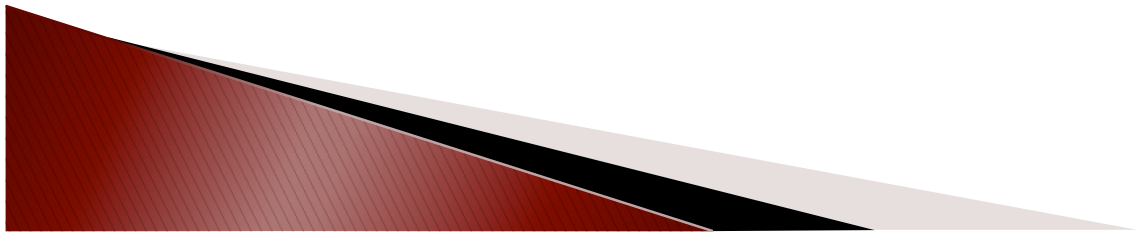
First thing's first

- What's the only thing you should believe?



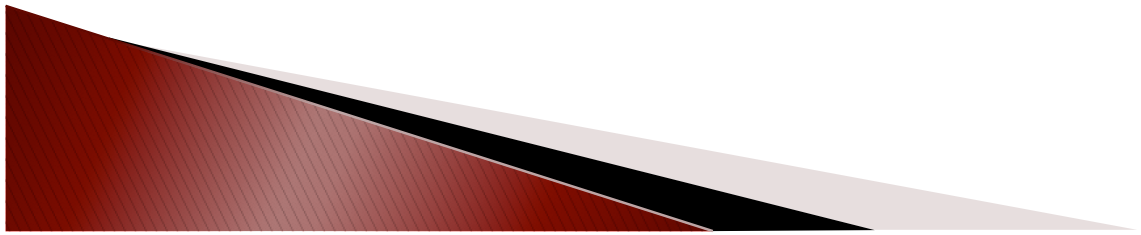
Purpose

- Ever checked something in to your repository with statements like “this need to be fixed” knowing damn well you never will go back and fix it?
- Ever write something random in the commit statements because you know no one will ever read them?
- Ever invent new profanity in comments?
- Yeah, we’ve done all that too
- This project’s goal is to analyze commit statements and comments looking for amusing, evil, and interesting things



Code Repository Primer

- No, wait…



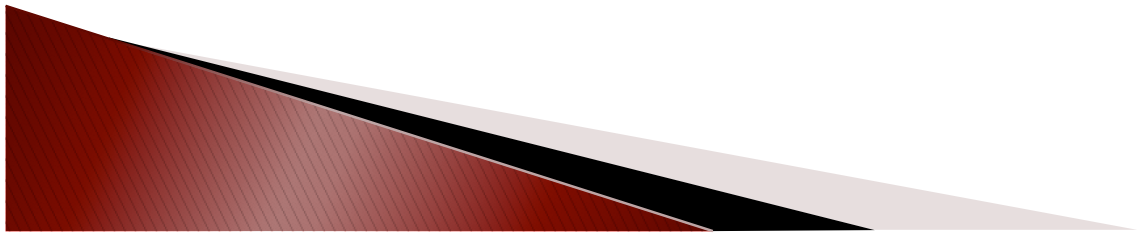
A Primer on "Primer"

- A primer (in this case) is loosely defined as "a description of elementary issues for a given topic"
- It is pronounced **prim**-er
- Primer (**prahy**-mer) is the shit you put on walls before you paint

- Want an example? Watch Jodie Foster in "Contact"
  - Then watch her in Taxi Driver… just cuz

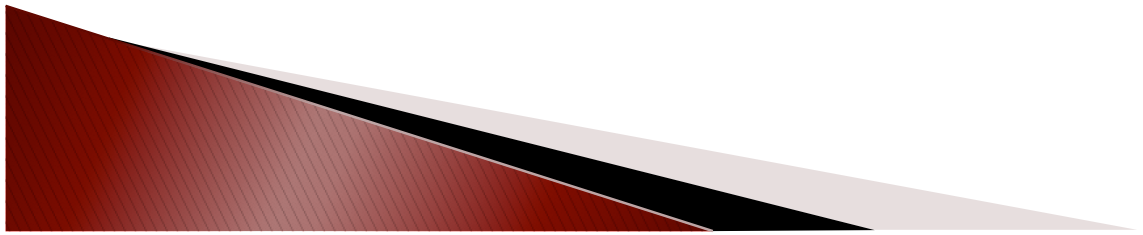
Back to Repositories

- Three major source repository software suites
  - CVS
  - SVN
  - GIT
- They're all impressively different
  - There's also terabytes of publicly available repositories
- Yeah.. There are other repos. You can debate each over in the corner with the GNU/Linux folks
- We focused on GIT and SVN
  - SVN because we know it
  - GIT because you can get everything via "git clone".



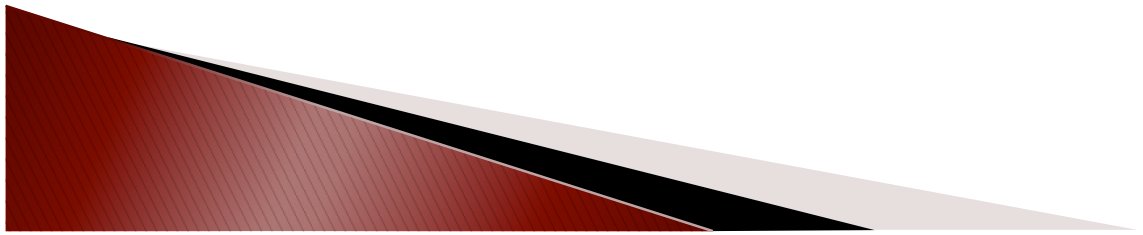
What projects did we examine

- Mostly C-based repositories from places like
  - Github
  - Sourceforge
  - And just googling for strings that indicate a public repos
- Some python and others
- It's a surprise



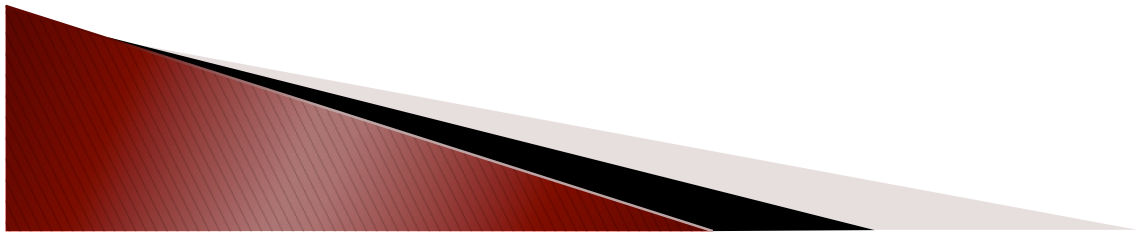
Methodology

- Identify targets
  - SF/Github/Google searches
- Take URL list and go git them (hahaha)
- Shove files in to a hashed directory structure
  - Needed a way to quickly navigate a HUGE file store (stat() kinda sucks)
  - MD5 hash of the name, then broke down directories based on hash value
  - Could have stored it in a DB, but honestly that would be grotesque



Methodology

- Parse all the flat files that we got from the repos and put interesting things in the MySQL database
  - Comments
  - Commit statements
  - Other stuff
- Create a web interface to allow for robust querying
  - Turbogears makes source comments sexy

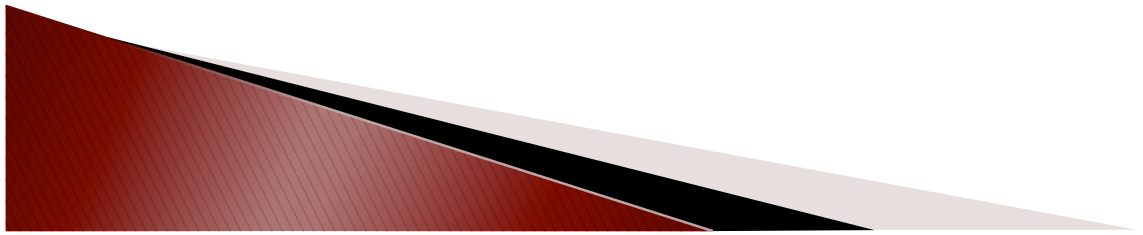


Crawler details

- Python-based
- Must be aware of what's been crawled already
  - Saves bandwidth
  - Keeps us under the radar
    - What we're doing arguably is frowned upon by many ToS for the sites we're dealing with
- Holy disk usage
  - We definitely made a space vs. time trade off with the crawler…
  - Use up more disk space on the front end to save on processing time later

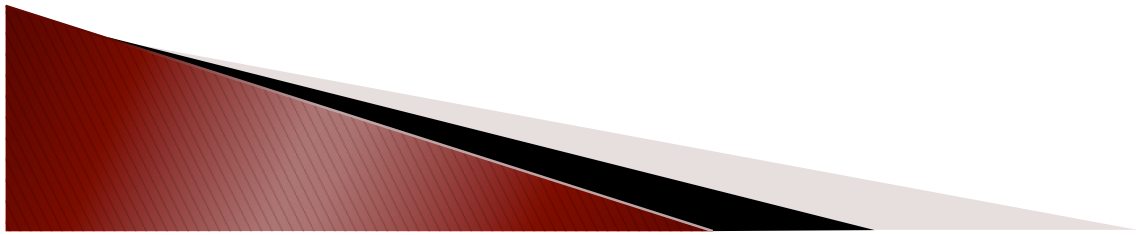
Parsing / Analysis

- Initially simple string searching (grrrrrep)
  - Dirty words
  - Specific words such as “security”, “needs to be fixed” ,etc
  - Interesting results
- Moving on to more sophisticated grammatical analysis



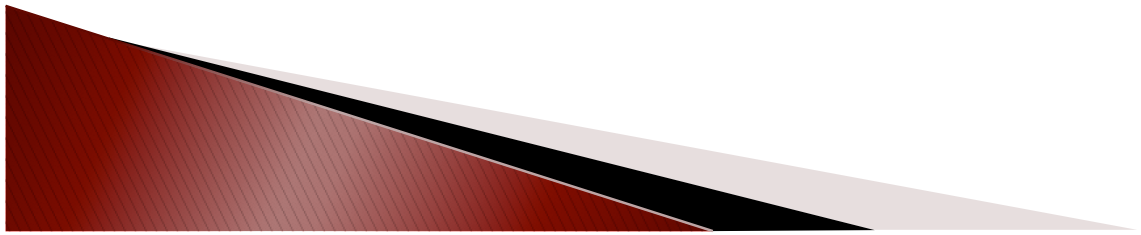
Demo

- Live demo at Defcon



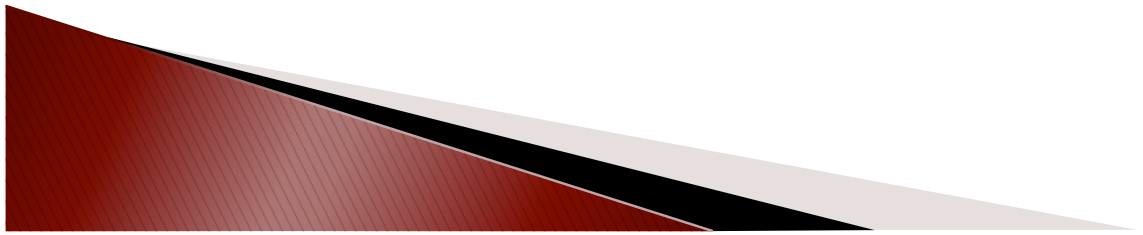
Results – Amusing

- TBD



Results – Whitespace only

- Ever just hit space in the commit statement?
  - Yeah, others have as well



Results – Offensive

- TBD

Results – Arguments

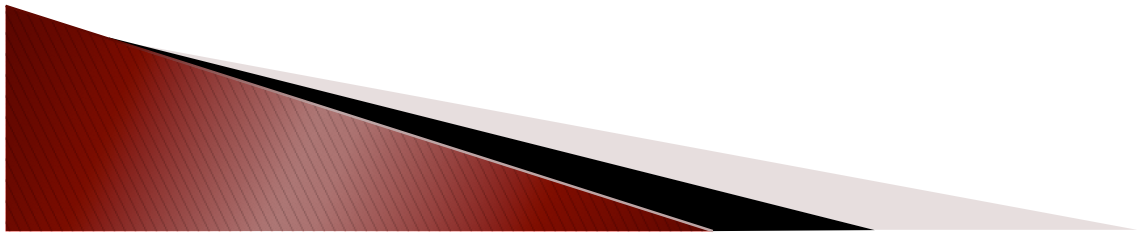
- TBD

Results – Social commentary

- TBD

Results – Security Specific

- TBD



Conclusion

- What you way when you commit sticks with you and your project forever
- There's a TON of source code on the net
  - This is trivial analysis
  - Think about doing actual static code analysis at scale…. Right.

- This presentation is for the DefCon CD and will be updated before the con. For a complete version, go to www.shmoo.com