Mining Evolution Data of a Product Family

Michael Fischer, Johann Oberleitner, Jacek Ratzinger

Vienna University of Technology

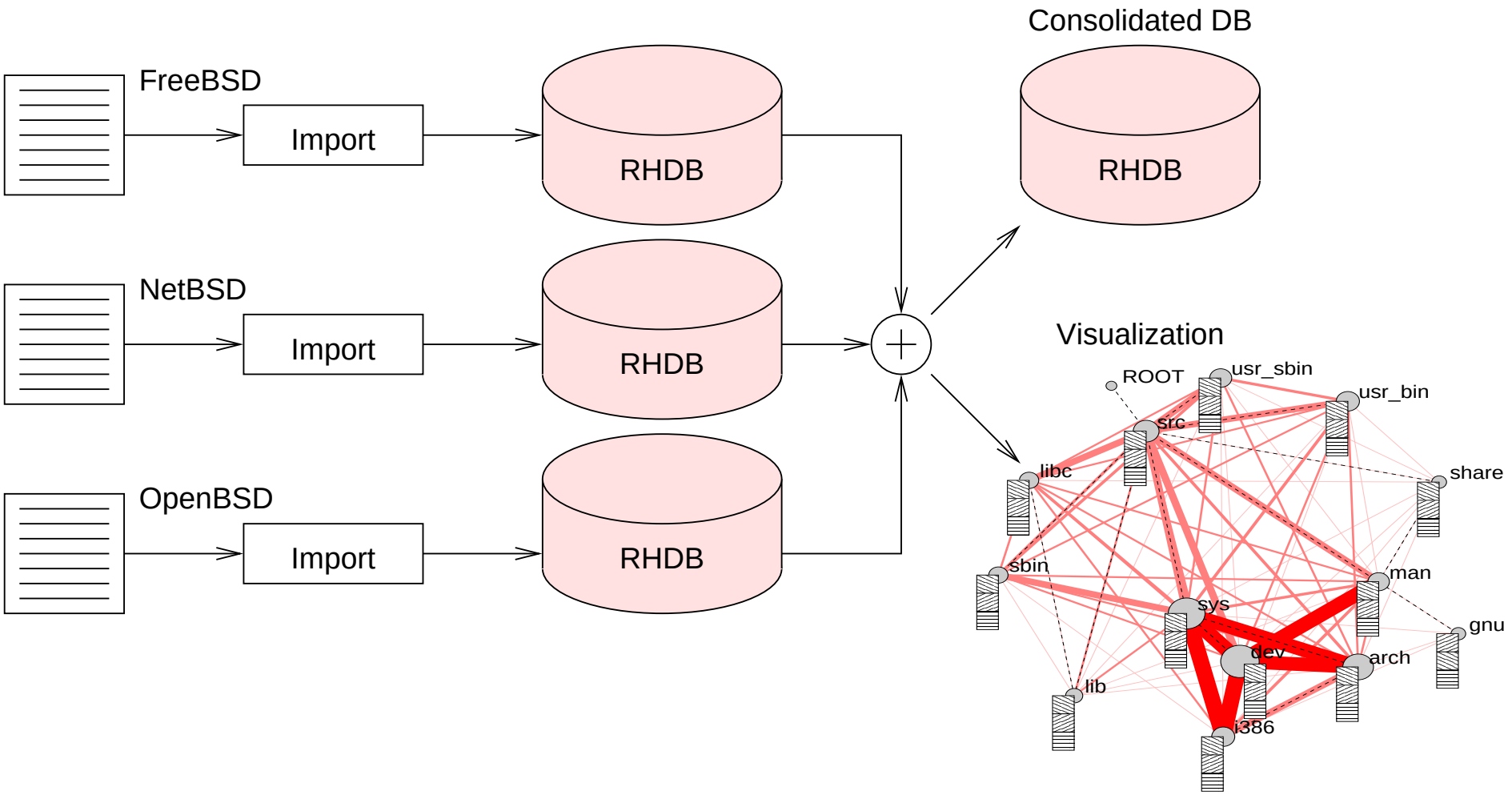
Harald Gall

University of Zurich

Product Family Evolution

- Objective: find relationships between variants of a product line
- Approach:
  - Use lexical search in change log messages
  - Retrieve change dependencies from RHDB
  - Visualize findings about change dependencies and information flow between source code directories via Multidimensional Scaling (MDS)
- Case study:
  - Product variants: FreeBSD, NetBSD, OpenBSD (30.000 - 60.000 files)
  - Direct copies of CVS systems
  - Used keywords: *freebsd*, *netbsd*, *linux*

Process


FreeBSD
Import
RHDB
NetBSD
Import
RHDB
OpenBSD
Import
RHDB
Consolidated DB
RHDB
Visualization
ROOT
usr_sbin
usr_bin
src
libc
share
sbin
man
sys
gnu
dev
arch
lib
i386

Common files

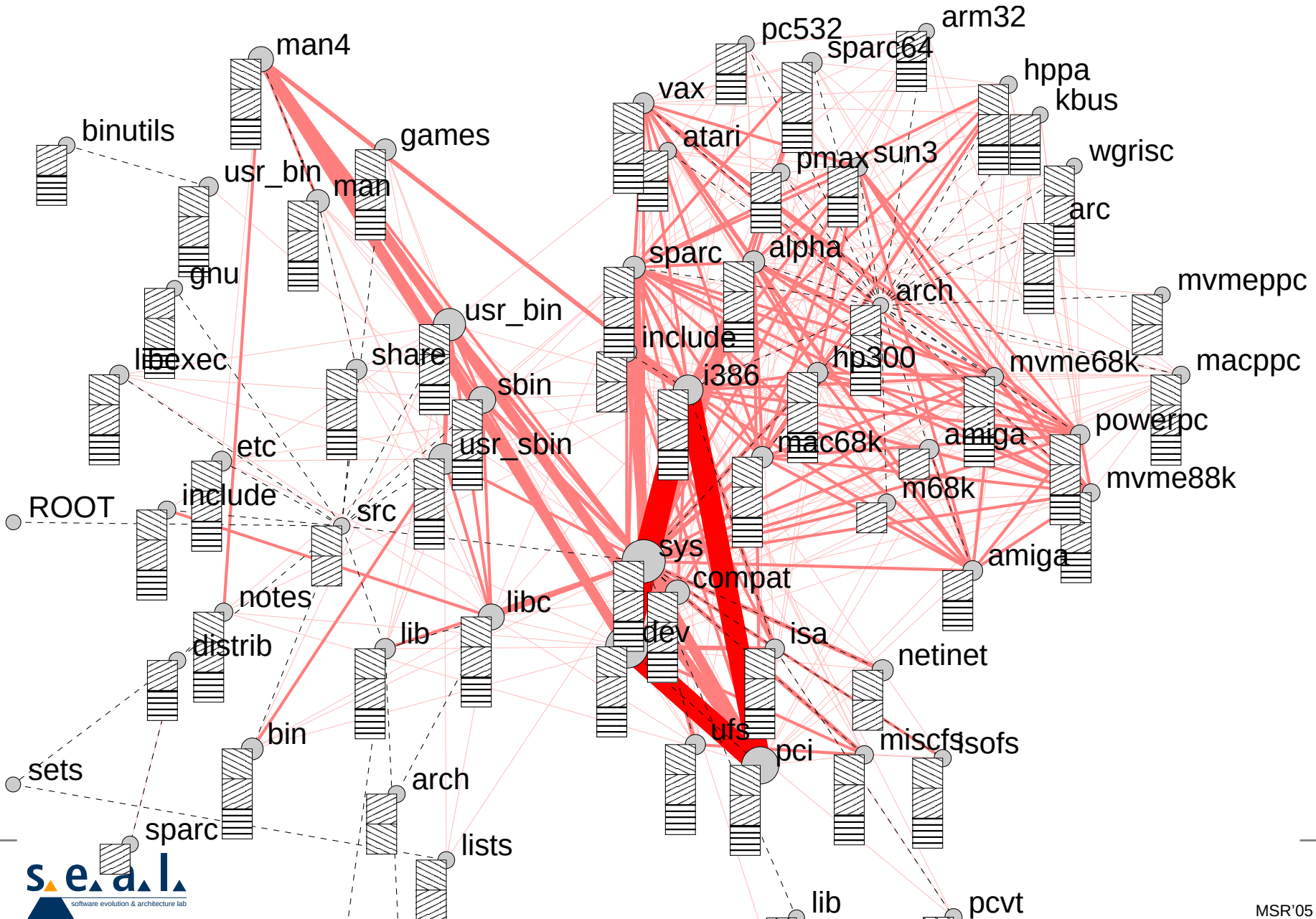
Variant	Variant	all modules	*src/sys/* only
FreeBSD	*NetBSD*	3810	1333
FreeBSD	*OpenBSD*	3839	1079
NetBSD	*OpenBSD*	6969	6847

Information flow

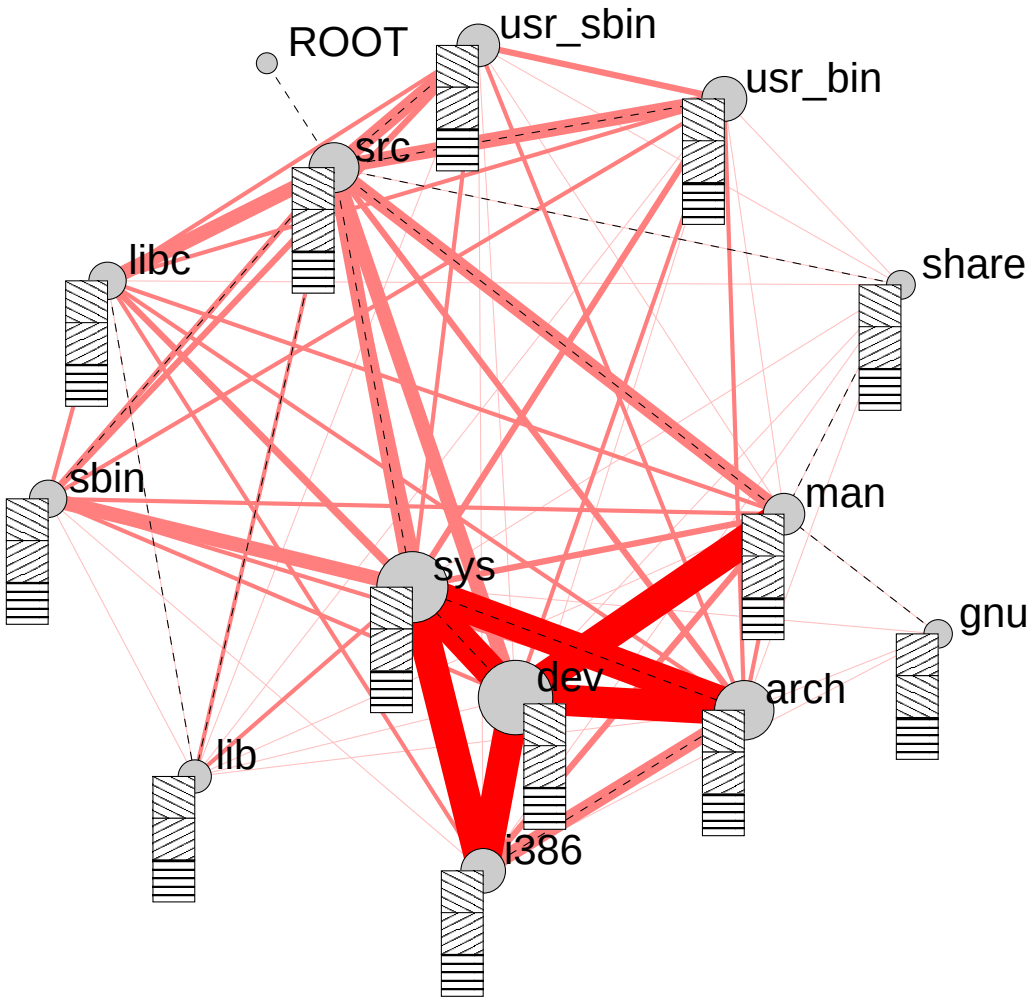
- Information flow between variants of the BSD systems based on lexical search

Variant	Keyword	all revisions	revision > 1.1
FreeBSD	*netbsd*	5131	3577
	openbsd	2729	1353
	linux	1791	1387
NetBSD	*freebsd*	2852	2186
	openbsd	2679	2224
	linux	1547	1125
OpenBSD	*freebsd*	2406	1933
	netbsd	16802	7423
	linux	775	463

Global view: *OpenBSD*



Abstracted view: *OpenBSD*



Referenced files: *OpenBSD*

- Topmost referenced files with one of the given keywords in the change logs of *OpenBSD*

Keyword	Count	Path
freebsd	59	src/sys/dev/pci/files.pci
.	52	src/sys/dev/pci/pciide.c
.	52	src/sys/dev/pci/pcidevs
netbsd	45	src/sys/arch/i386/i386/machdep.c
.	43	src/sys/dev/pci/pciide.c
.	39	src/sys/conf/files
linux	14	src/sys/compat/linux/linux_socket.c
.	14	src/sys/compat/linux/syscalls.master
.	5	src/sys/dev/ic/if_wireg.h

Detailed change analysis

- Change in *FreeBSD* file *ufs_quota.c*

```
< sleep((caddr_t)dq, PINOD+2);
---
> (void) tsleep((caddr_t)dq, PINOD+2, "dqsync", 0);
```

- Change in *NetBSD* file *ufs_quota.c* (six years later)
- Change in *OpenBSD* file *ufs_quota.c* (eight years later)

```
< sleep((caddr_t)dq, PINOD+2);
---
> (void) tsleep(dq, PINOD+2, "dqsync", 0);
```

Conclusions & Possible Extension

- Lexical search indicates increasing information flow between product variants
- Recovered high level view indicates high coupling and wide spread of “alien” source code
- Text mining in modification reports
  - Vector space search (latent semantic indexing)
- Code clone detection
  - Clone detection in source code
  - Track clone propagation through version history
  - Clustering
    - *Measuring Similarity of Large Software Systems Based on Source Code Correspondence* [Yamamoto et al.]
    - *CCFinder* [Kamiya et al.]