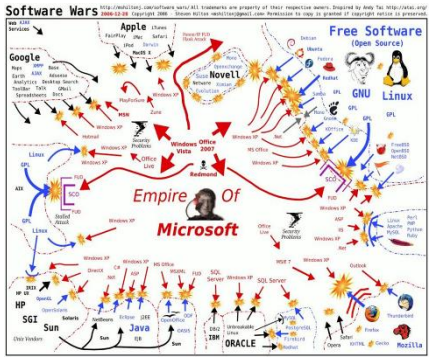
开源数据分析

周明辉

数据驱动的~开源软件及其社区/生态的~度量、分析、预测和智能化支持



开源大数据

海量成功案例和最佳实践

开源开发复杂系统的机制机理、方法技术和支撑工具

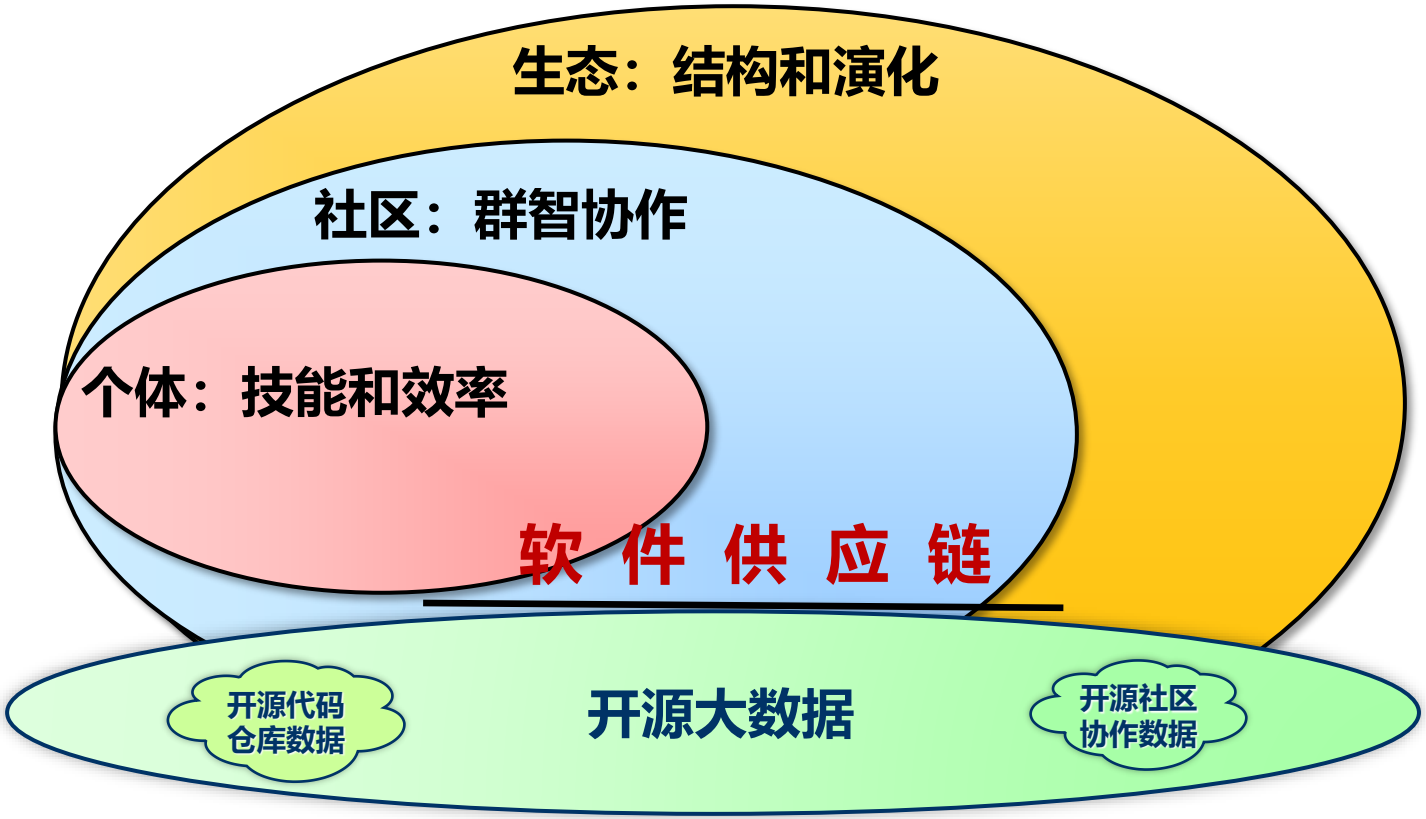
复杂sys原理和技术

开源数字社会学/开源动力学

北京大学开源分析实验室：
https://osslab-pku.github.io

开源数字社会学

- 研究基础：开源大数据
- 研究方法：数据驱动方法：数据挖掘、智能分析、学科交叉的方法和洞见
- 研究对象：开源代码/软件 ~ 开发个体/群体
  - 发现型研究：设计精细度量仪，观测和度量复杂系统，发现规律
  - 发明型研究：设计智能系统&自动化bot，支持自动化软件开发和群体协作



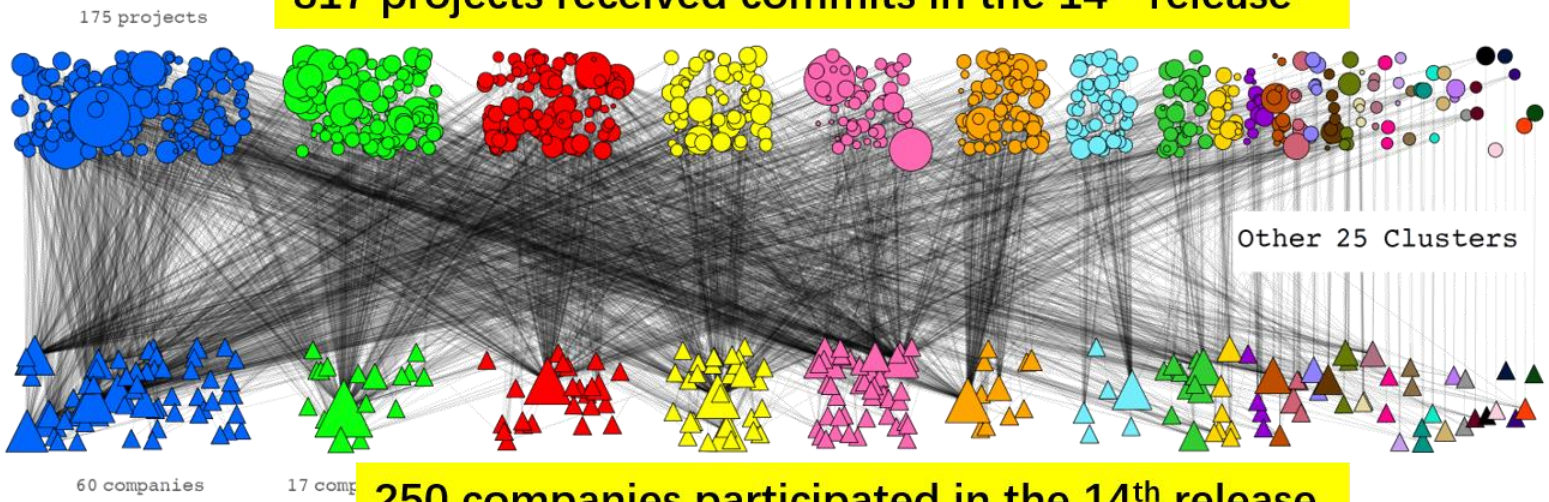
发现型研究：精细度量，发现规律

- 开源贡献者的技能&行为度量
  - 全网技能名片，初始行为决定未来贡献
- 开源社区&生态的要素、结构和法则
  - 复杂系统的结构刻画
- 开源软件供应链的要素、结构和演化
  - 风险传播，网络脆弱性分析

复杂生态的精细度量

OpenStack 14th : 817code repo/34,192patch/2,439dvpr/250org

817 projects received commits in the 14th release



250 companies participated in the 14th release

Three Collaboration Patterns:

- Intentional
- Passive
- Isolated

Intentional collaboration:
- Supply and Consumption
- Distribution-oriented ally
- Service delegation

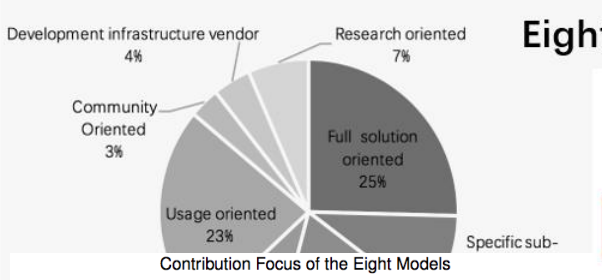
Rack space — Ansible — Walmart

Passive collaboration:
- Out of own interest
- Most common

HP — Magnum — CERN

Isolated collaboration:
- Plugins/drivers
- New services

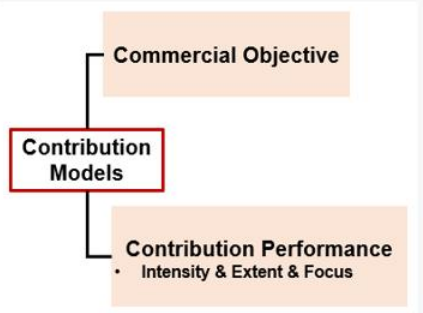
Citrix — fuel-plugin-xenserver



Contribution Focus of the Eight Models

Contribution Model	Contribution Focus
Full solution oriented	Computing
Specific sub-solution oriented	Storage
Self-business oriented	Networking
Specific services oriented	Documents
Usage oriented	Deployment
Community oriented	Infrastructure of development
Development infrastructure vendors	Infrastructure of development
Research oriented	Orchestration

Eight Contribution Models



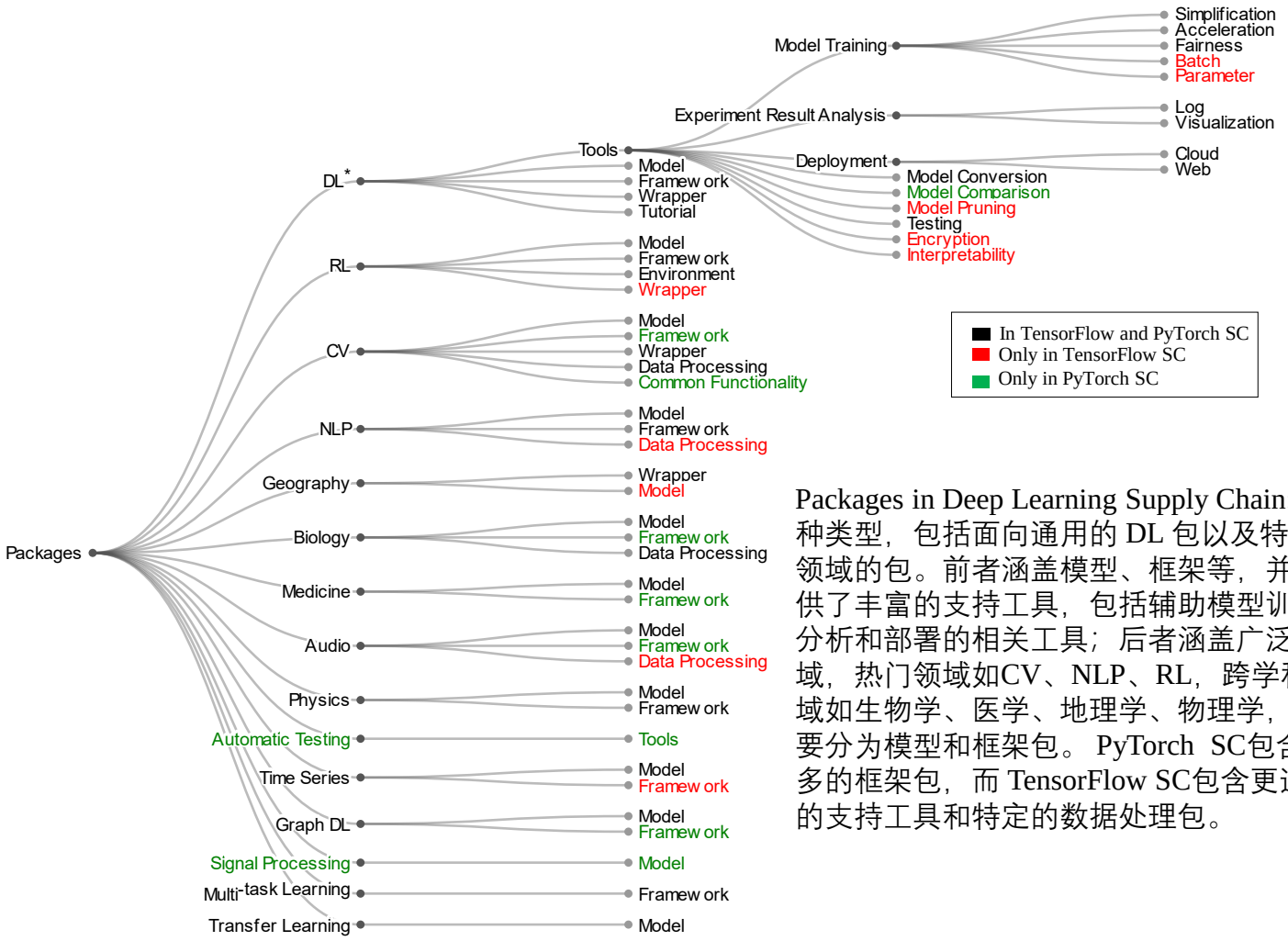
Four Types of Task Preference:

- Correction Focused
- Feature Focused
- Reengineering Focused
- Mainly Correction and Feature

Zhang et al. Companies'participation in oss development -- an empirical study of openstack. IEEE Transactions on Software Engineering, 2019.
Zhang et al. How do companies collaborate in open source ecosystems? ICSE'2020
Zhang et al. Corporate dominance in open source ecosystems: a case study of OpenStack. FSE'2022

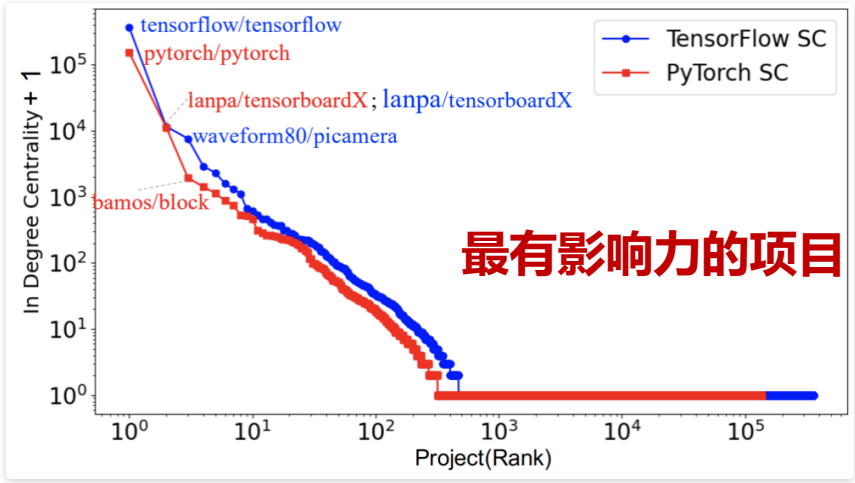
AI供应链的网络结构、关键节点和风险传播

- AI供应链：以TensorFlow和PyTorch为基础建立project/package依赖链条
- Data: World Of Code + Libraries.io

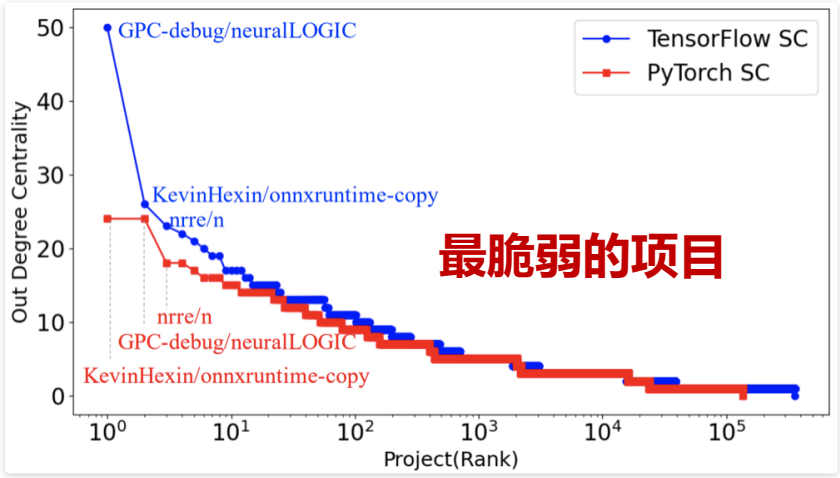


Packages in Deep Learning Supply Chain: 52种类型，包括面向通用的 DL 包以及特定领域的包。前者涵盖模型、框架等，并提供了丰富的支持工具，包括辅助模型训练、分析和部署的相关工具；后者涵盖广泛领域，热门领域如CV、NLP、RL，跨学科领域如生物学、医学、地理学、物理学，主要分为模型和框架包。 PyTorch SC包含更多的框架包，而 TensorFlow SC包含更通用的支持工具和特定的数据处理包。

Xin Tan et al. An Exploratory Study of Deep Learning Supply Chain. ICSE'2022



In-degree Distribution of the Projects.



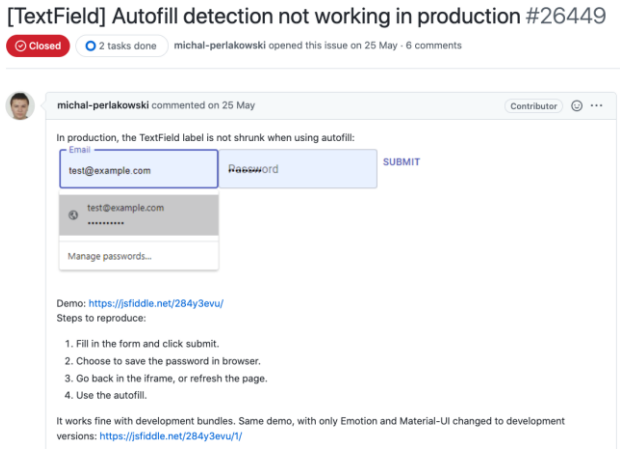
Out-degree Distribution of the Projects.

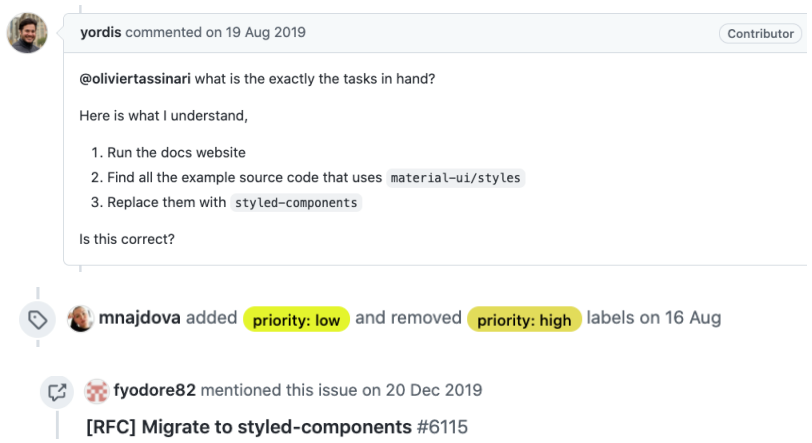
发明型研究：建立智能系统，自动推荐

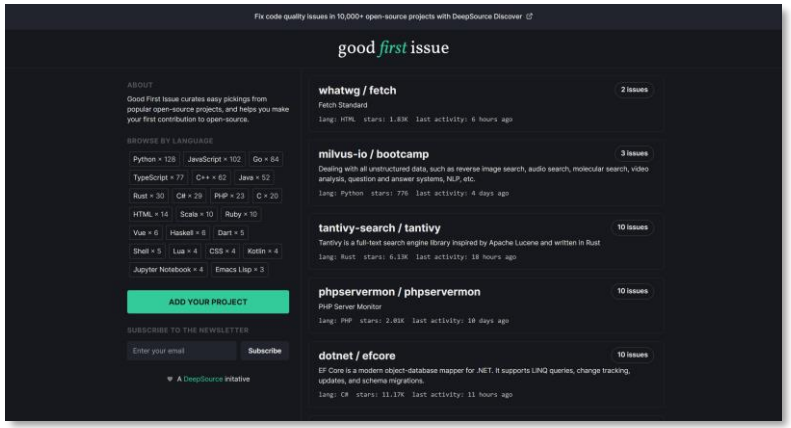
- 写代码：API recommendation
  - 场景：软件需要将其使用的第三方库迁移到另一个功能相同或类似的第三方库。
  - 输出：利用数据挖掘技术构造有效量度，推荐迁移替代库。
- 写注释：Inconsistent comment and code detection
  - 场景：注释解释代码，但其经常过期，因此需要检测不一致注释以做修正。
  - 输出：挖掘大规模代码文件，提取特征，检测/修正失配的代码注释。
- 开源社区支持：自动识别项目中新手友好任务
- 开源程序员画像：基于全网数据的程序员开源名片

开源项目新手任务推荐：GFI Recommendation

- AI-Powered智能推荐: 为开源项目issue/任务自动打标签，即判断其是否“简单独立”任务。



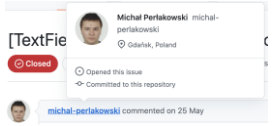




Issue的标题和描述: 是否清晰，是否提供了足够的信息。

Issue的标签:中低难度类, Doc类更可能是GFI;
Difficult/Important, Coding, Build类最不可能是GFI。

Issue是否有项目维护者的参与: 讨论、打标签、提及等。



背景信息: issue报告者的经验。
在该项目中具有少量经验的开发者提出的issue更可能是GFI。

https://github.com/osslab-pku/gfi-bot

模型(*f*): 回归模型(XGBoost等)。*p*=*f*(*x*)
训练集: 历史issue数据,令被新人(解决issue时在项目没有贡献过commit)解决的issue的*p*=1,其余issue的*p*=0。
测试点: 任一未被分配的open issue。

输出(*p*): 一个issue是GFI的概率。

He et al. GFI-Bot: Automated Good First Issue Recommendation on GitHub. ESEC/FSE’2022
Xiao et al. Recommending good first issues in GitHub OSS projects. ICSE’2022
Tan, Zhou and Sun. A First Look at Good First Issues on GitHub. FSE’2020

发现型研究：大规模的库迁移推荐实证研究

- How frequently do projects migrate a library?
  - 8.98% to 28.72% projects have undergone at least one library migration
  - For those projects, most have no more than five migrations.
- How do migrations happen between libraries?
  - Library migrations from four domains (logging, testing, JSON, and web service) among 53 dominate the dataset, presenting a long tail distribution.
  - Library migrations are highly unidirectional in that most libraries are either mostly adopted or mostly abandoned.

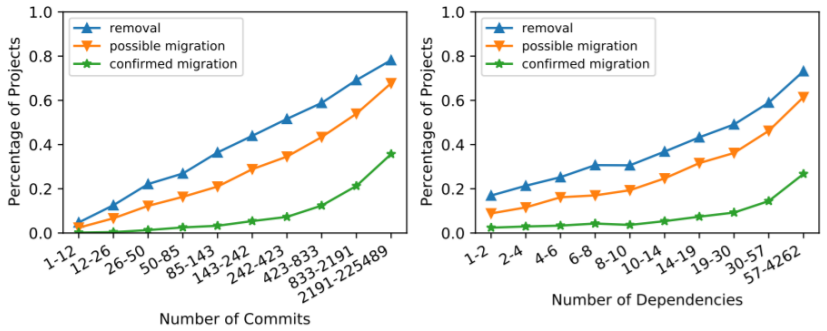


Figure 4: Distribution of P_{cm} and P_{pm} by number of commits and dependencies. We also show the results of $P'_r = \{p | p \in P_m \wedge D_p^- \neq \emptyset\}$ for comparison.

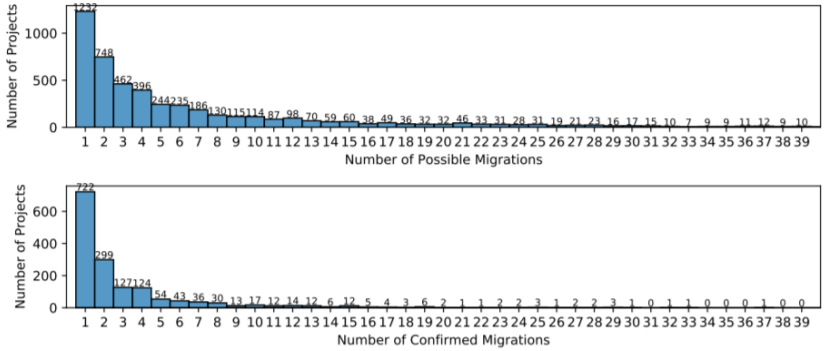


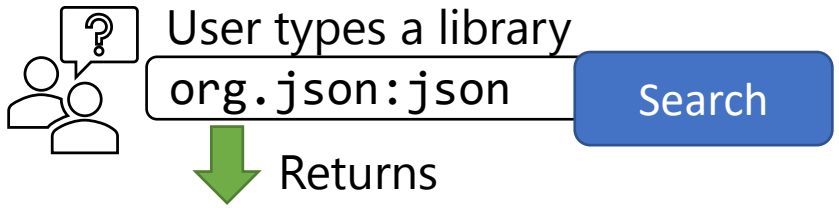
Figure 5: Distribution of projects by number of migrations.

ESEC/FSE 2021

Hao He, Runzhi He, Haiqiao Gu, and Minghui Zhou. A large-scale empirical study on Java library migrations: prevalence, trends, and rationales. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (*ESEC/FSE 2021*).

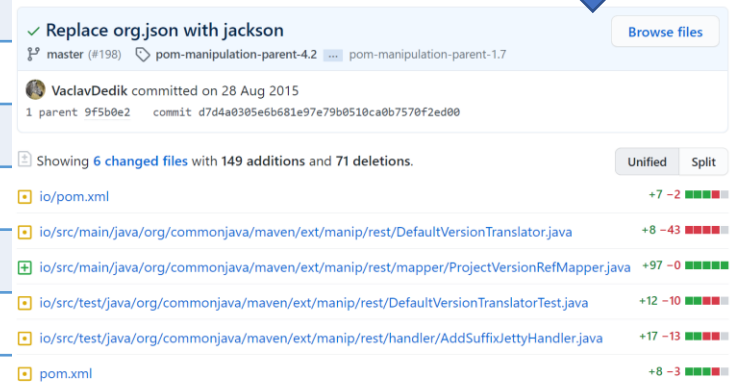
发明型研究：库替换推荐工具

- 在线工具可公开访问: http://migration-helper.net/
- 数据集可公开访问: https://zenodo.org/record/5091384
- 已作为华为开源治理服务IDEA插件部署



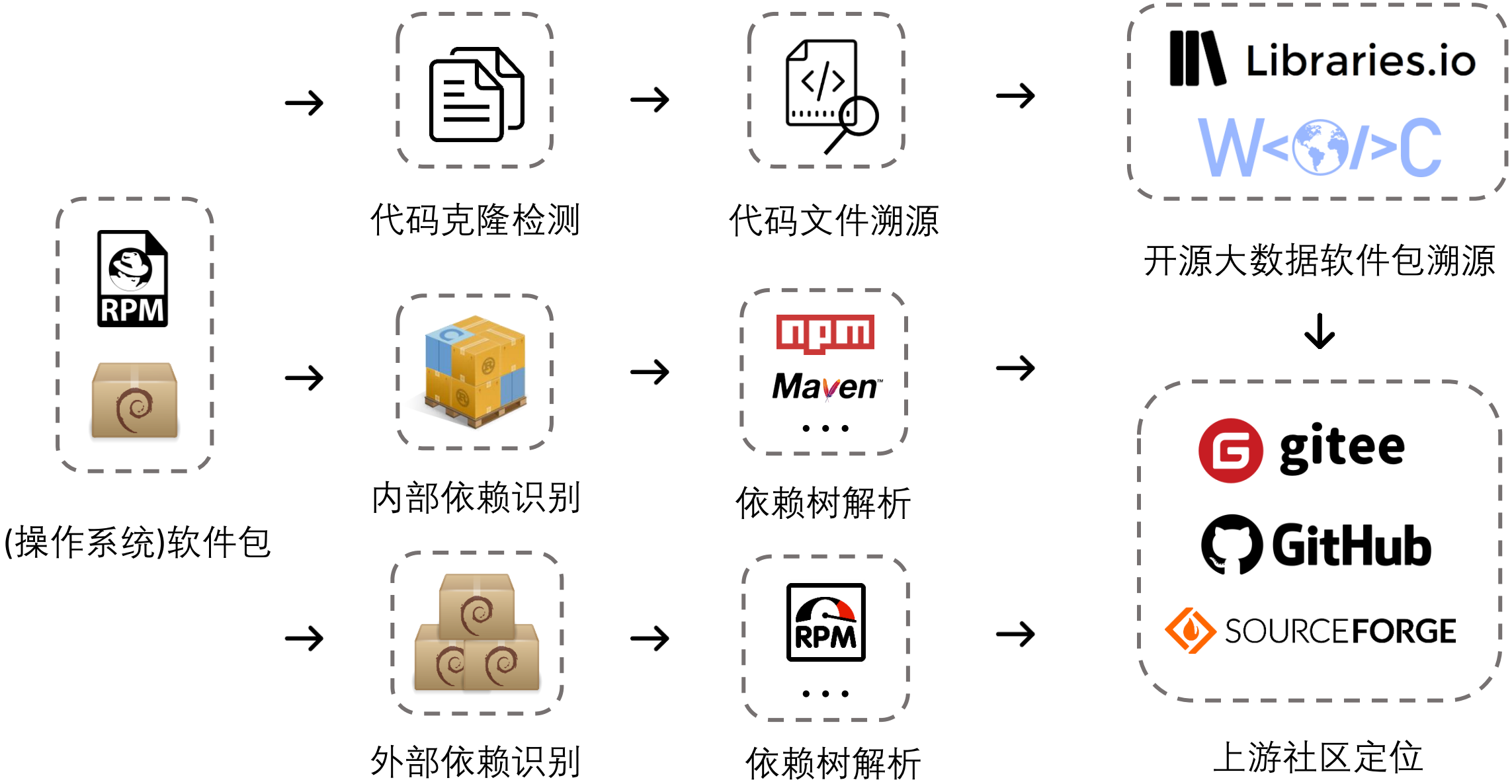
Rank	Target Library	Confid
1	com.fasterxml.jackson.core:jackson-databind	
2	com.google.code.gson:gson	
3	com.alibaba:fastjson	
4	org.apache.felix:org.apache.felix.utils	
5	com.googlecode.json-simple:json-simple	
6	org.apache.johnzon:johnzon-core	
7	com.fasterxml.jackson.core:jackson-core	
8	org.springframework.boot:spring-boot-starter-json	

Repository	Commit
apache/asterixdb	54ecc63
release-engineering /pom-manipulation-ext	0f2ec00
opendaylight/aaa	315b73
.....	.....

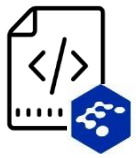


H. He, Y. Xu, X. Cheng, G. Liang and M. Zhou, "MigrationAdvisor: Recommending Library Migrations from Large-Scale Open-Source Data," *2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, 2021, pp. 9-12,

开源软件供应链构建：方法



开源软件供应链构建：目标

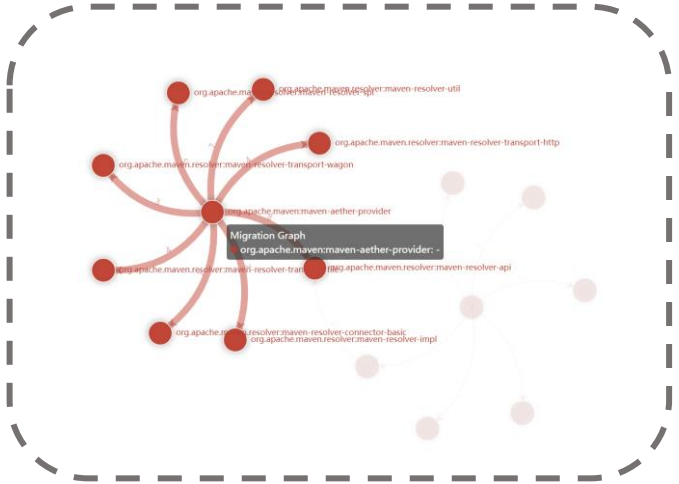


扫描OS源代码

Libraries.io

W</>C

软件包溯源



可视化供应链查询工具

+

Bill of Materials

Name	Version	Security Score	Match Level	Path	Number of CVEs
ben	0.6.1ubuntu2	100	Low	Demo/Example/COTS.zip/libtidy.dll	0
expat	2.0.1	100	High	Demo/Example/COTS.zip/program.exe	0
freetype	2.8.1	35	High	Demo/Example/COTS.zip/CoreGraphics.dll	1
libcue	2.2.1	100	Low	Demo/Example/COTS.zip/CoreGraphics.dll	0
libflac	1.3.2	100	Low	Demo/Example/COTS.zip/CoreAudioToolbox.dll	0
libjpeg	6b2	29	High	Demo/Example/COTS.zip/CoreGraphics.dll	2
libnvidia-gl	460.67	100	Low	Demo/Example/COTS.zip/Admin.dll	0
libnvidia-gl	460.67	100	Low	Demo/Example/COTS.zip/program.exe	0
libpng	1.2.56	2	Low	Demo/Example/COTS.zip/gnsdk_manager.dll	7
libpng	1.2.56	2	Low	Demo/Example/COTS.zip/CoreMedia.dll	7
libpng	1.2.56	2	Medium	Demo/Example/COTS.zip/CoreGraphics.dll	7
libtiff	3.9.4	9	Medium	Demo/Example/COTS.zip/CoreGraphics.dll	55
libxml	2.9.3	100	High	Demo/Example/COTS.zip/WebKit.dll	0
libxml	2.9.4	100	High	Demo/Example/COTS.zip/libxml2.dll	0
libxslt	1.1.28	2	High	Demo/Example/COTS.zip/libxslt.dll	10

可集成复用的供应链知识库（SBOM）

Collaboration on Open Source Practice in China: 平台/社区、协议、基金会、项目

- 中文开源平台的实现与支持
  - GitLink | 确实开源
  - 木兰开源社区 (mulanos.cn)
- 木兰开源许可证及许可证推荐:
  - http://license.coscl.org.cn/MulanPSL2
  - https://licenserec.com
- 企业开源治理：开源软件供应链，开源社区要素度量
- 开源项目： World-of-Code, GFI-Bot, MigrationAdvisor, LicenseRec…

End