

From Developer Networks to Verified Communities: A Fine-Grained Approach

Mitchell Joblin Wolfgang Mauerer Sven Apel, Janet Siegmund Dirk Riehle
ICSE '15 Proceedings of the 37th International Conference on Software
Engineering - Volume 1 p563-573

背景

- ソフトウェア工学はソフトウェアシステムを構成するための協調作業である
- ソフトウェアシステムはサブコンポーネントごとに分割し、管理されている
- 開発者はサブコンポーネントの依存関係を考慮した上で、他の開発者と作業する必要がある

先行研究

- 開発者のコミュニティがソフトウェアの品質に重大な影響を与える
- 組織の特徴に基づいた評価方法の方が、ソースコードに基づいた方法より欠陥を指摘する
- 開発者のネットワークを構築するためにVCS(version control system)を用いられてきている

先行研究と本研究の目的

- 先行研究
 - スモールワールド現象をみたすようなネットワークの特性を特徴づけることを主な焦点としている
- 本研究の目的
 - よりきめの細かい組織化(サブコミュニティ)をした特徴を明らかにすることを焦点としたネットワーク構築手法を提案する

アプローチ

- VCSからデータを取得
- 仮説を立てて、既存の手法と本研究で提案する手法を比較する
- 仮説をもとに開発者にアンケートを行い提案手法の適合率を確認する

既存手法と提案手法

- 既存手法

- A. ファイルベースでの共同作業を考慮した手法

- B. コミットのsign-offタグを用いた手法

- 提案手法

- A. 機能ベースでの共同作業を考慮した手法

- B. committerとauthorの関係を考慮した手法

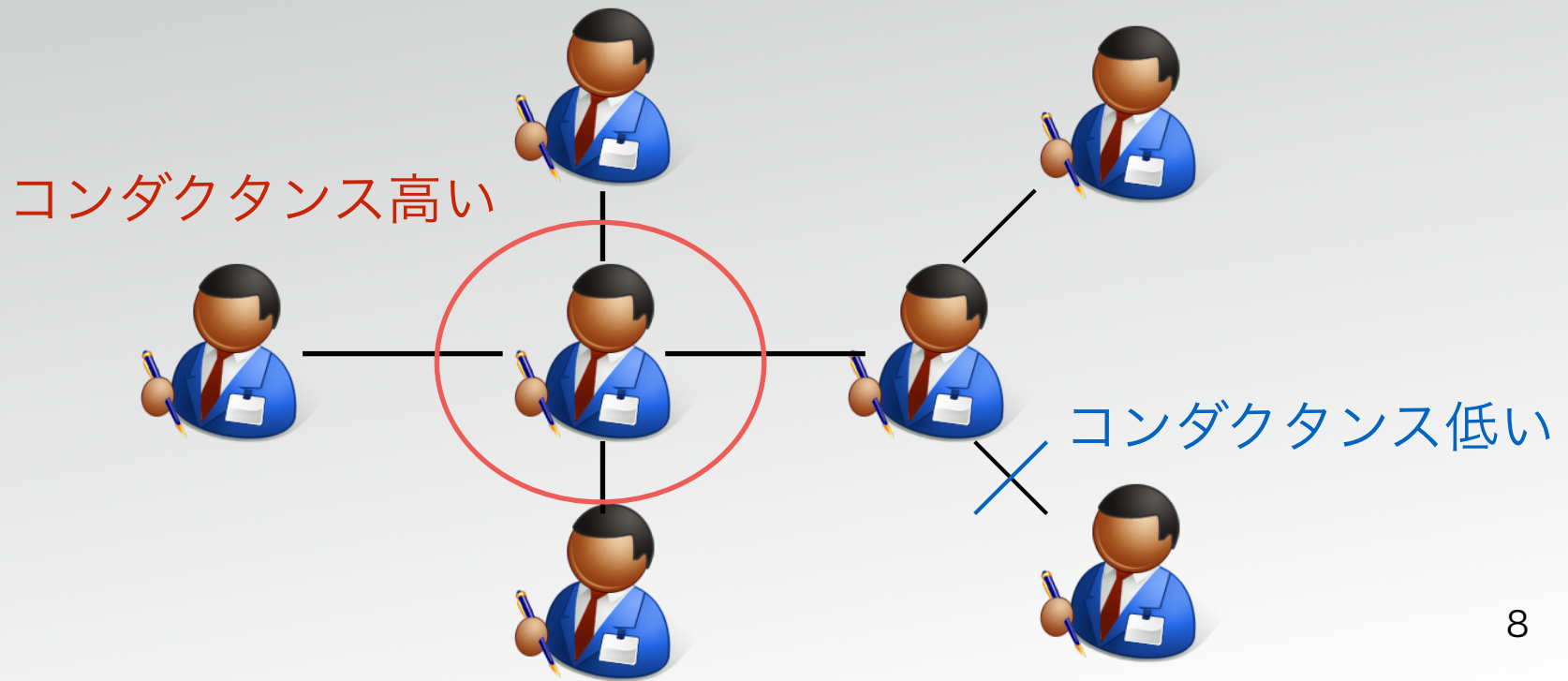
分析手法

- 開発者ネットワークの分析
 - A. 重み付き有向グラフを用いた検出アルゴリズム (OSLOM)の適用(部分一致も検出可能)
 - B. ランダムグラフを用いたnullモデルと提案手法を用いたモデルをコンダクタンスを用いて比較

コンダクタンス

- ノードの集合 V から部分集合 v_1 を切り取るときの各エッジの重みの総和

- コミュニティのネットワークの場合、値が高いとそのノードは隣接するノードと濃い関係であるといえる



対象プロジェクト

Project	Devs	MLOC	Lang	Domain
Linux	580	16	C	OS
Chromium	500	6.5	C, C++	User
Firefox	400	9.3	C++, JS	User
GCC	70	6.2	C, C++	Devel
QEMU	50	0.78	C	OS
PHP	50	2.2	PHP, C	Devel
Joomla	30	1.3	PHP, JS	Devel
Perl	30	4.5	Perl, C	Devel
Apache http	15	2.2	C	Server
jQuery	15	0.05	JS	Library

仮説

- H1
 - 開発者ネットワークは無作為に作成された予測されていない組織の重要性を超える同一のコミュニティを示す
- H2
 - ファイルベース手法を用いたネットワーク構造は重要なコミュニティを確認できず、機能ベース手法を用いた手法は重要なコミュニティを確認できる
- H3
 - タグベース手法を用いたネットワーク構造はcommitter-authorを用いたネットワーク構造と一致する

H1:結果

(開発者ネットワークは無作為に作成された予測されていない組織の重要さを超える同一のコミュニティを示す)

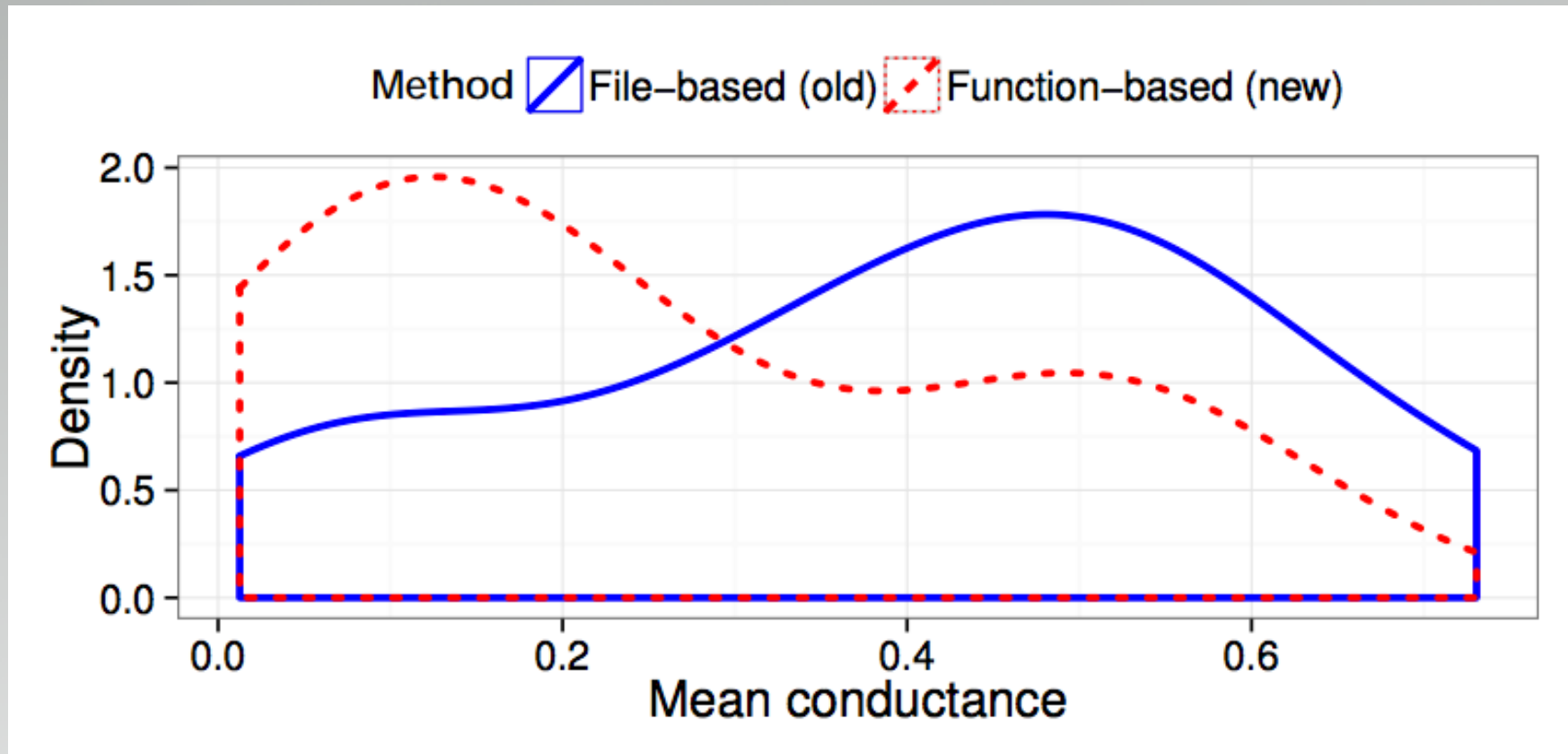
TABLE I

OVERVIEW OF SUBJECT PROJECTS FOR A 90-DAY DEVELOPMENT WINDOW AND COMPARISON OF COMMUNITY CONDUCTANCE FOR THE ORIGINAL (OBSERVED) AND RANDOMIZED (REWired) NETWORKS.

Project	Devs	MLOC	Lang	Domain	Observed conductance	Rewired conductance
Linux	580	16	C	OS	0.05	0.88
Chromium	500	6.5	C, C++	User	0.20	0.74
Firefox	400	9.3	C++, JS	User	0.11	0.79
GCC	70	6.2	C, C++	Devel	0.01	0.48
QEMU	50	0.78	C	OS	0.39	0.56
PHP	50	2.2	PHP, C	Devel	0.15	0.80
Joomla	30	1.3	PHP, JS	Devel	0.57	0.84
Perl	30	4.5	Perl, C	Devel	0.49	0.66
Apache http	15	2.2	C	Server	0.27	0.80
jQuery	15	0.05	JS	Library	0.49	0.75

結果:H2

(機能ベースの方が重大なコミュニティを特定できるか)



提案手法の方が
高密度かつ強固なコミュニティを構築することができた

例:QEMUのコミュニティ

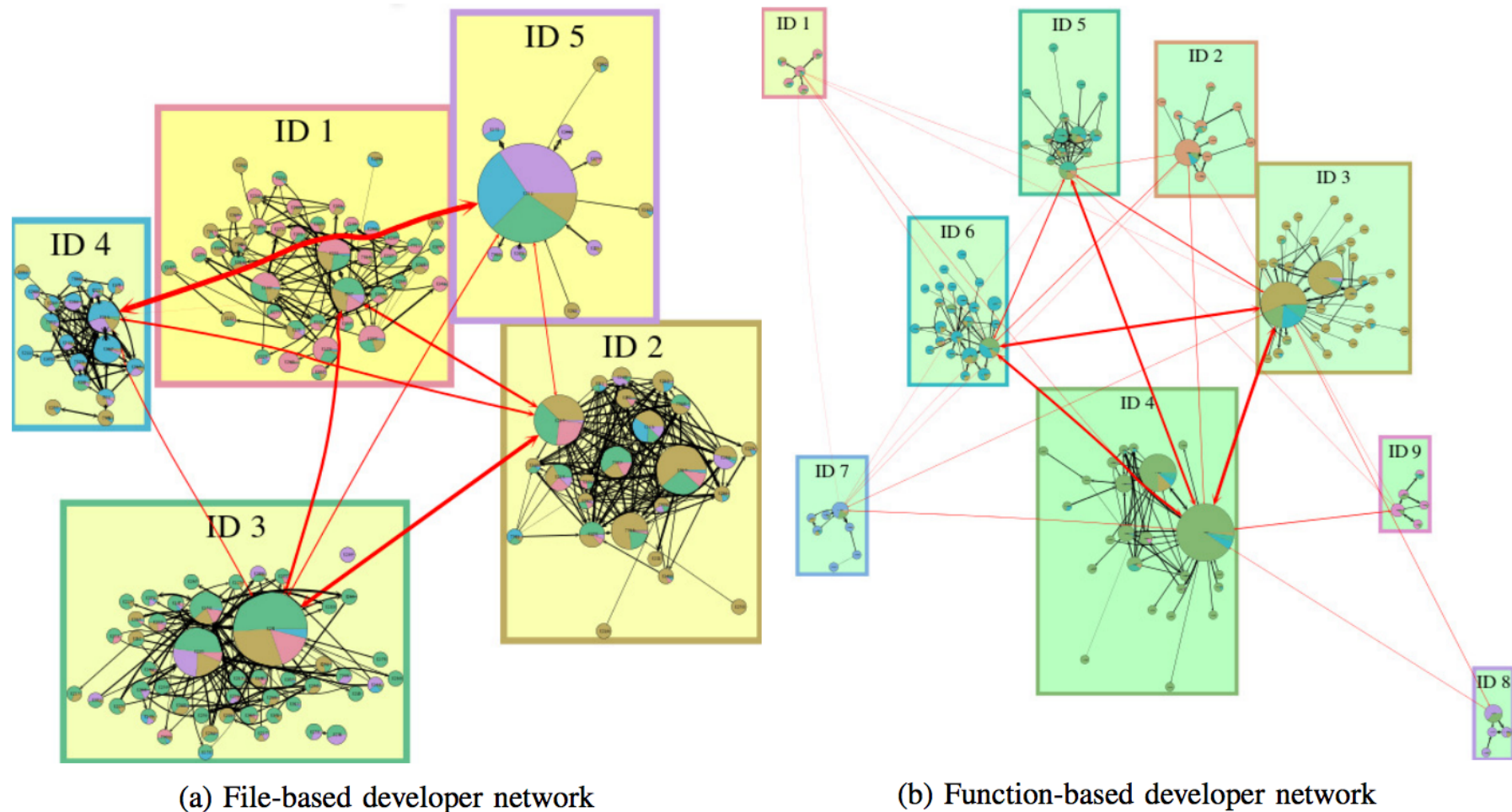
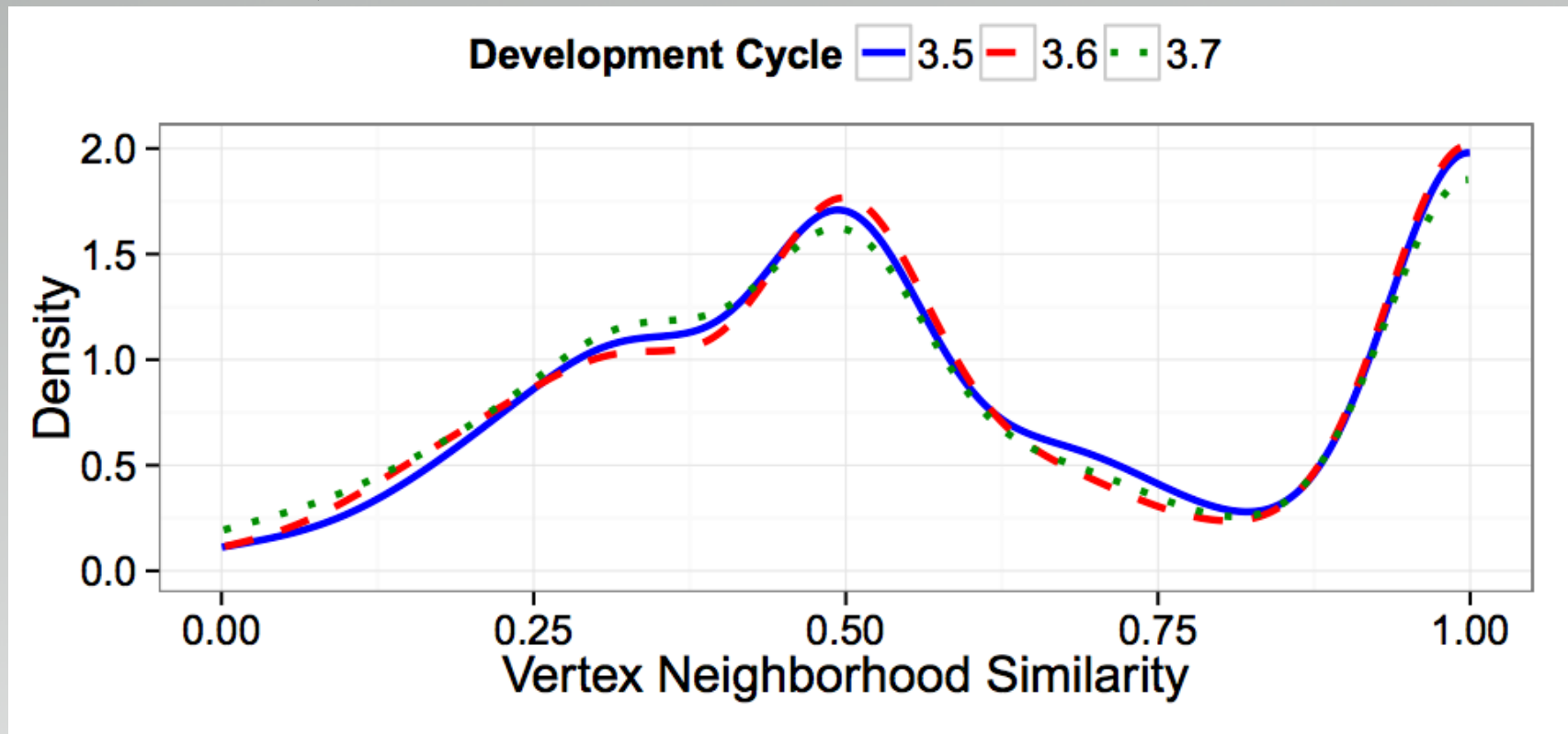


Fig. 4: Developer networks constructed from QEMU v1.1. All communities in the file-based network are insignificant (yellow background color). The function-based method identified several significant communities (green background color).

結果:H3

(タグベースとコミッター・オーサー間の関係ベースは一致するか)…例:Linux



全体で70%一致した

リサーチクエスチョン

- RQ1
 - 構築したネットワークは実際の間敬意を性格に表しているのか？
- RQ2
 - 同一と確認の取れたコミュニティは実際の重要性おを表しているのか？

リサーチクエスチョンのアプローチ

- アプローチ
 - 各プロジェクトの開発者にアンケートをとる
 - ランダムに選出した521人中53人の返答があった
- Survey Question
 - SQ1:バージョンXにおいて誰とよく協力したか？
 - SQ2:ネットワークは開発者の正しい関係を表しているか？
 - SQ3:ネットワーク上に示された開発者はコミュニティに存在しているか？

回答:SQ2,3

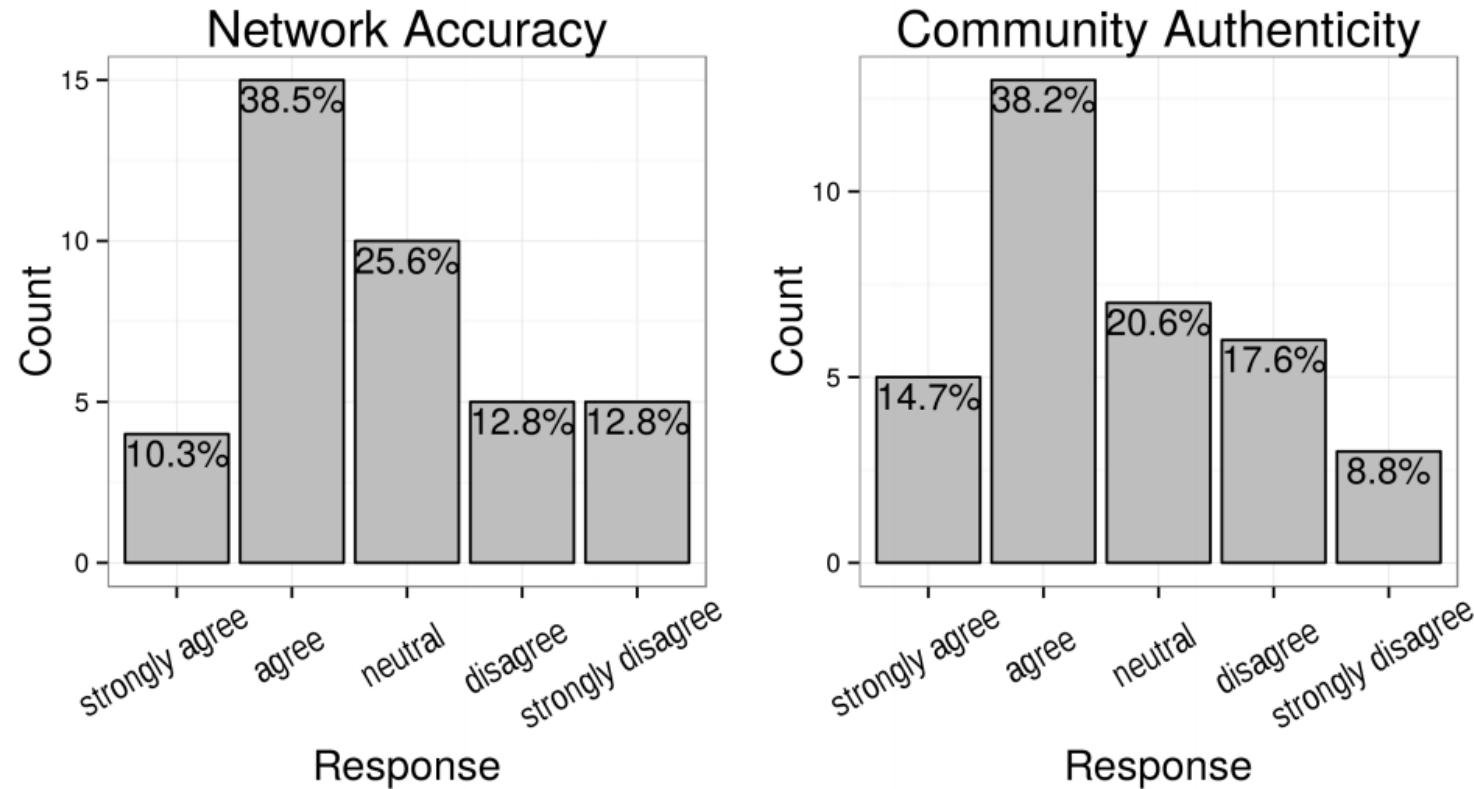


Fig. 6: Developer survey responses to questions stated in Section IV-F (SQ2 left, SQ3 right).

多くの意見で賛同を得られたが、ノードの連結・コミュニティの構成に誤りも見られた

機能ベース手法とファイルベース手法 の比較

- ファイルベース手法で現れたノードから機能ベース手法で現れたノードを除きSQLの回答をもとにファイルベース方式の整合性を確かめる
- 結果として、ファイルベース手法は機能ベース手法よりも大きいコミュニティを形成するため、多くのノードが見られたが、その約80%は誤ったものであった

妥当性の考慮

- 対象のプロジェクトが10と少ないので一般的であるとは言い難い
- タグベースネットワークは、サインオフタグを使用しているプロジェクトが少ないため一般的であるとは言い難い
- 商用プロジェクトの場合ははっきりとしていない

まとめ

- 機能ベース，コミッターオーサー間の関係を基準にしたコミュニティ構成を明らかにする手法を提案した
- 先行研究ではできなかったコミュニティ構成を明らかにし可視化することができた
- しかし，妥当性に関しては考慮する必要がある

所感

- 提案手法は自分の研究にも使えるかも
- tag-basedが一例のみなのでどうなんだろう？