

Mining Email Social Networks

Christian Bird, Alex Gourley,
Prem Devanbu, Michael Gertz
Anand Swaminathan

目次

- 背景・目的
- 前処理
- ソーシャルネットワークの分析
- まとめ

背景・目的

背景

- 大規模開発では開発者間の交流は成果物と互いに影響を及ぼしあっている
- 開発者間の交流を分析して、その性質やそれが成果物にどのように影響を与えているかを知りたい

対象となるデータ

- 商用のソフトウェア開発での開発者間の交流はあらゆる事情があり、扱えない
- 自由に扱えるOSSのメーリングリストとCVSの履歴を扱う

ソーシャルネットワーク

- メーリングリストの中で、メールの送信や返信がどのように交わされているかを表すネットワーク
- 本論文であらゆる分析を行う際に用いられる

目的

- ソーシャルネットワークの性質を調べる
- 開発者間の交流と開発者の活動の関係を調べる
- 開発者と非開発者の役割の違いを調べる

前处理

ソーシャルネットワーク 構成時の問題点

- 1人の開発者が複数のメールアドレスを所有することがある
- 複数のメールアドレスは同じ人物のものとなり、後の分析結果を誤ったものにしてしまう
- 同一人物の複数のメールアドレスをひとつのものとして扱う必要がある

IDの作成

- <name,email>という形のIDをメールアドレスから作成する

“Kojiro Noguchi” <KjrNog@wakayama.jp>



<Kojiro Noguchi , KjrNog@wekayama.jp>

クラスタリングアルゴリズム

- すべてのIDの間の 3 種類の類似度
 - 2つのIDのnameの類似度
 - 2つのIDのnameとemailの類似度
 - 2つのIDのemailの類似度

を計算する

- 上記の類似度のうち、最大のものを2つのIDの間の類似度とし、その類似度がしきい値を超えるものを同じクラスター（集団）に入れる

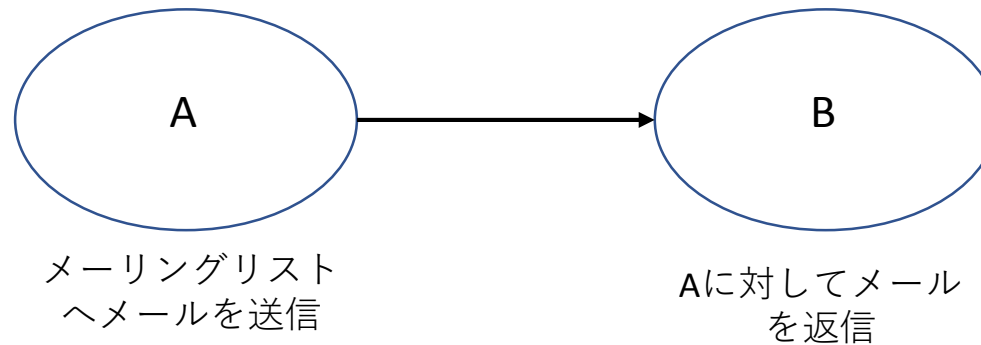
ソーシャルネットワークの分析

用いたデータ

- Apacheの1999年から2006年のメーリングリストの101,637件のメール
- その期間のApacheのCVSの履歴

ソーシャルネットワークの詳細

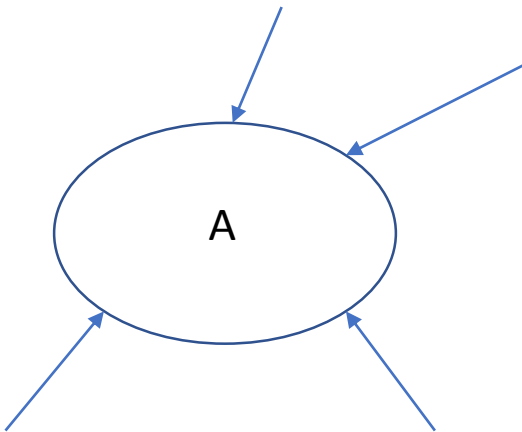
- ソーシャルネットワークは有向グラフ



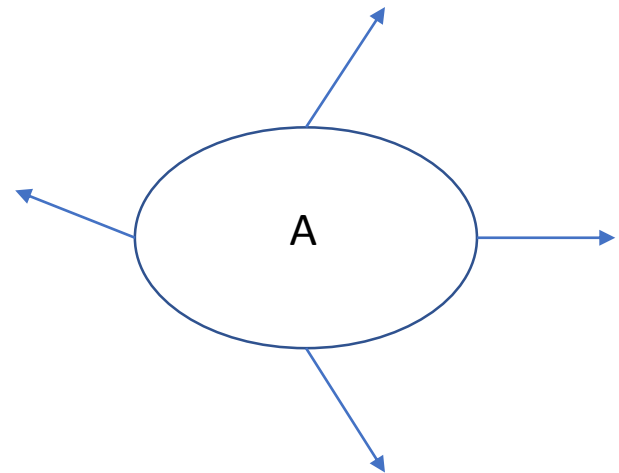
- 出次数が多いほど，その人がコミュニティの中で重要な人物であることを示す

入次数(in-degree)と出次数(out-degree)

- 入次数はメールを返信した数，出次数はメールを送信して返信された数を示す



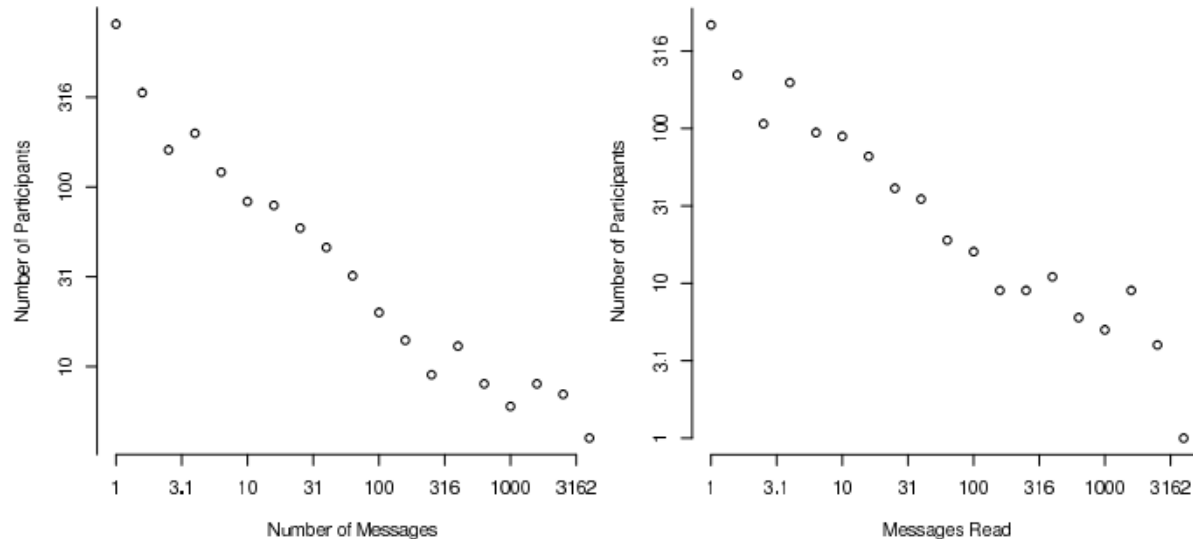
入次数



出次数

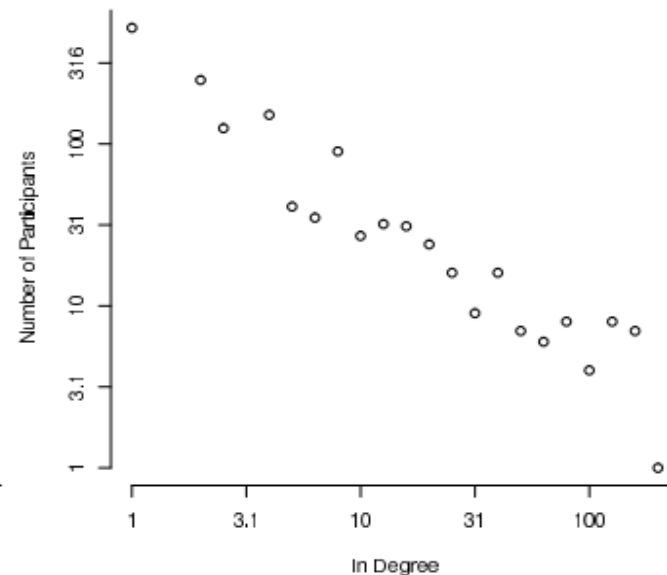
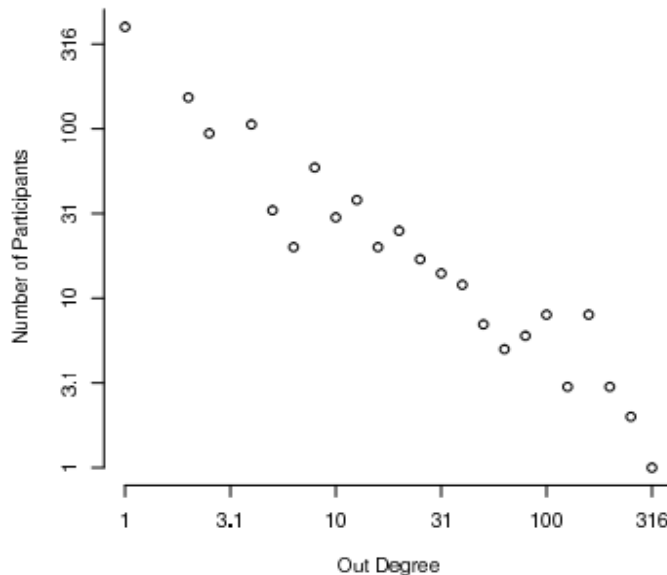
ソーシャルネットワークの性質

- 下のグラフの縦軸はどちらもMLの参加者数,
左のグラフの横軸はメール送信数, 右のグラフ
の横軸は返信数



ソーシャルネットワークの性質

- 下のグラフの縦軸はどちらもMLの参加者数, 左のグラフの横軸が出次数, 右のグラフの横軸が入次数
- 少数の人がコミュニティ内のかなり多くの人と交流していることが分かる



開発者間の交流と活動の関係

- ある開発者がどれだけ重要な役割を持っているかを知るために中心性を計算する
- グラフの最短路のうちノード v を通るのはどれだけあるかを示す

$$\text{中心性：} BW(v) = \sum_{i,j, i \neq j, i \neq v, j \neq u} \frac{g_{ivj}}{g_{ij}}$$

g_{ij} : 各最短路

g_{ivj} : 各最短路のうちノード v を通るもの

開発者間の交流と活動の関係

- ソースコードを書き換える役割を持つ開発者と入次数, 出次数, 中心性が大きい相関があることが分かる
- ソースコードを書き換える開発者はコミュニティ内の多くの人と交流していることが分かる

	changes	srcChanges	docChanges	outdegree	indegree	betweenness	mean	min	max
changes	1	0.789	0.932	0.520	0.474	0.553	912	0	16289
srcChanges	0.789	1	0.514	0.712	0.679	0.757	420	0	5741
docChanges	0.932	0.514	1	0.308	0.263	0.327	492	0	13420
outdegree	0.520	0.712	0.308	1	0.971	0.955	0.0080	0	0.0396
indegree	0.474	0.679	0.263	0.971	1	0.917	0.0067	0	0.0260
betweenness	0.553	0.757	0.327	0.955	0.917	1	0.0011	0	0.0965

開発者の活動とソーシャルネットワークの相関を示す図

開発者と非開発者の役割の違い

- 開発者と非開発者の中心性，出次数，入次数を比較すると，大きい差があることが分かる
- 開発者のほうが非開発者より重要な役割を担っていることが分かる

	Developer	Non- -Developer	T-value	Significance
Betweenness	0.0114	0.000140	5.07	$p < 0.001$
Out-degree	0.00666	0.000451	8.14	$p < 0.001$
In-degree	0.00794	0.000367	7.54	$p < 0.001$

開発者と非開発者の中心性，出次数，
入次数の違いを示す図

まとめ

- ソーシャルネットワークを見ると，小数の人がかなり多くの人と交流していることが分かった
- 開発者間の交流と活動にはかなり強い関係があることが分かった
- 開発者は非開発者に比べ，重要な役割を持っていることが分かった