Gender and Tenure Diversity in GitHub Teams

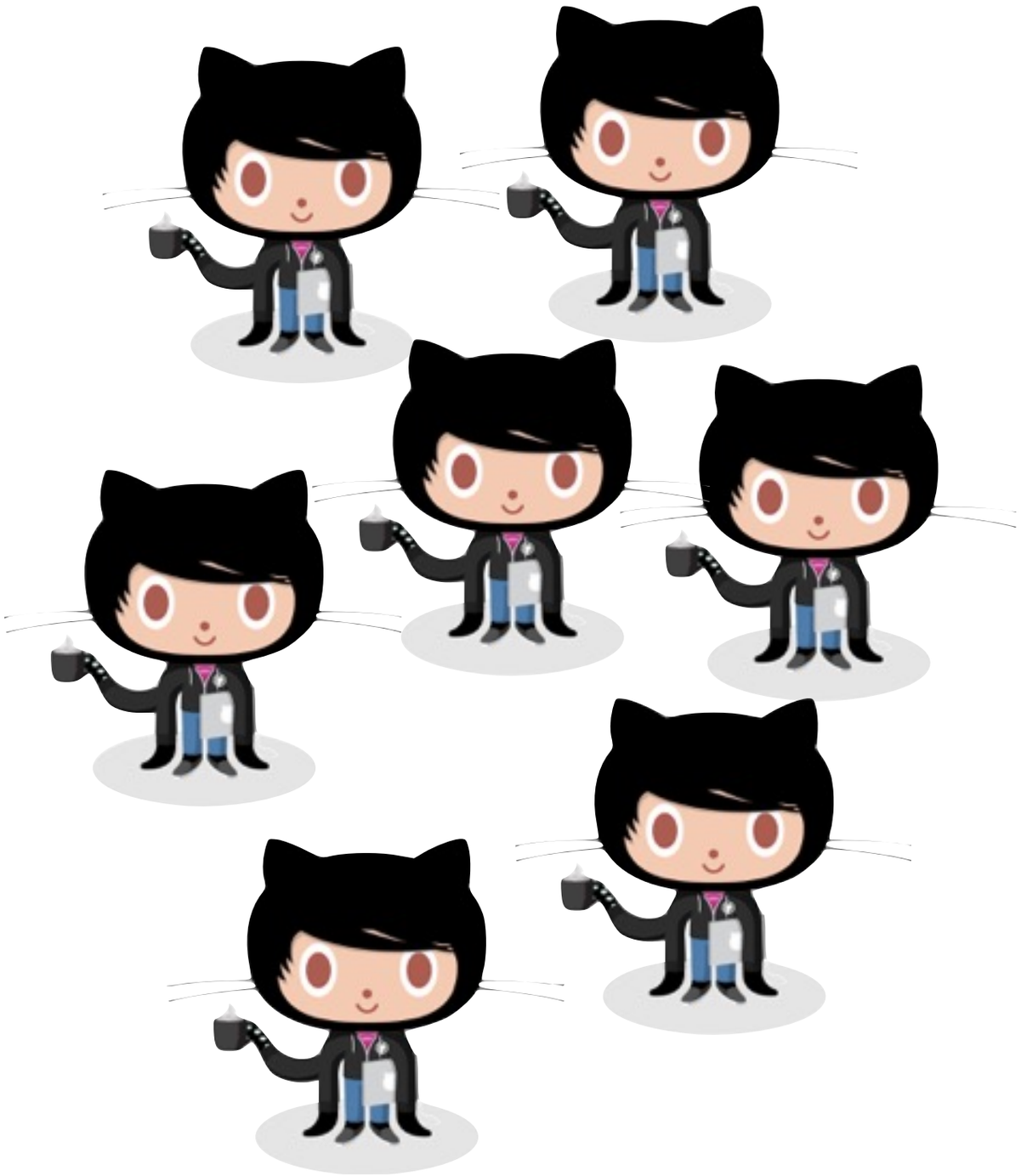
Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark van den Brand, Alexander Serebrenik, Prem Devanbu, Vladimir Filkov

UC DAVIS UNIVERSITY OF CALIFORNIA

TU/e Technische Universiteit Eindhoven University of Technology

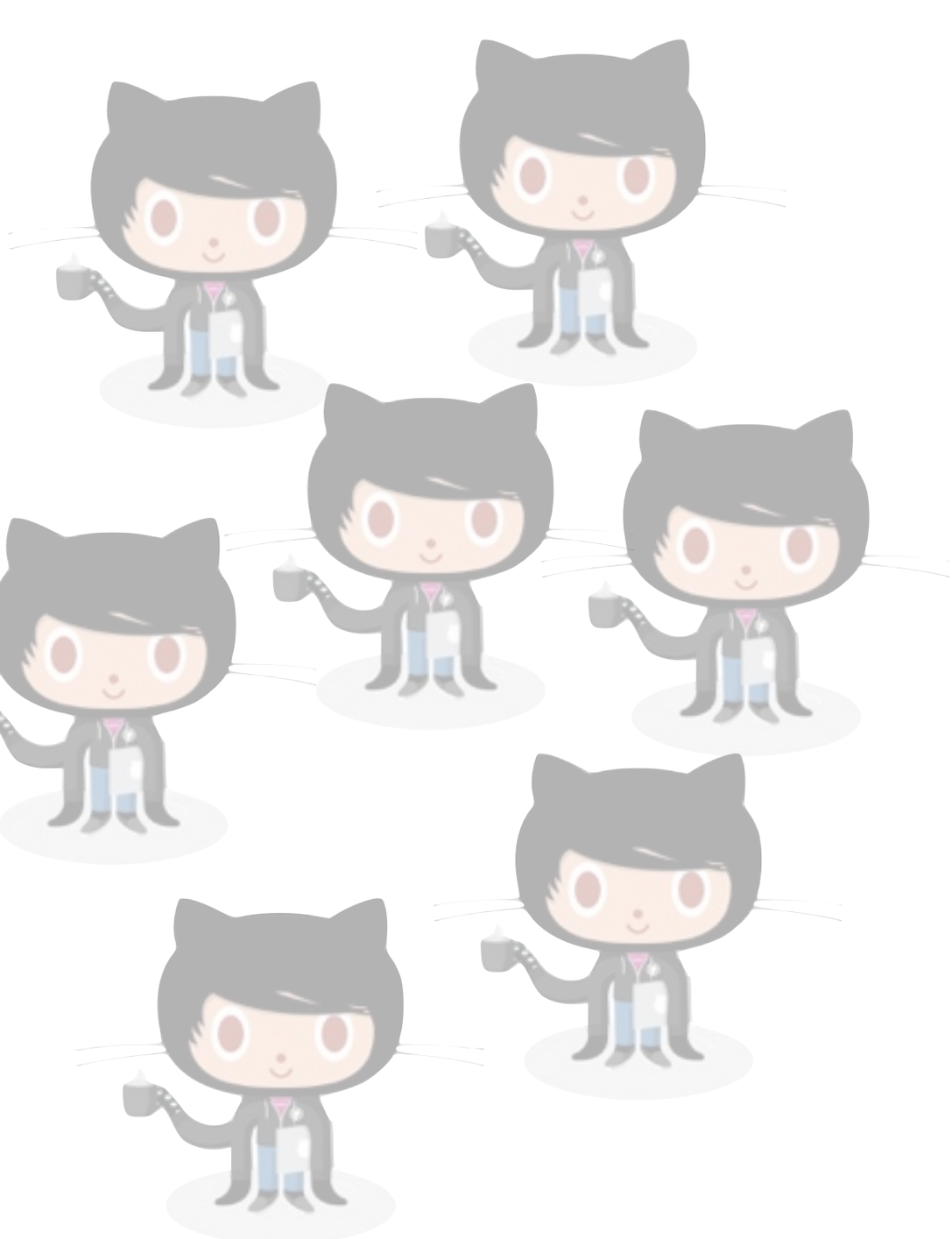
@b_vasilescu @baishakhir @MarkvandenBrand @aserebrenik @devanbu @vlfilkov

Which is more effective?



@b_vasilescu @baishakhir @MarkvandenBrand @aserebrenik @devanbu @vlfilkov

Which is more effective?



@b_vasilescu @baishakhir @MarkvandenBrand @aserebrenik @devanbu @vlfilkov

Diversity 👎

Similarity attraction theory

People prefer working with others similar to them in terms of values, beliefs, and attitudes [Byrne]

Social identity and social categorization theory

People categorize themselves into specific groups. Members of own group are treated better than outsiders [Tajfel]

Due to greater perceived differences between groups than within groups, diversity can lead to confusion, stress, and conflict [Horwitz & Horwitz]

Diversity 👍

Driver of internal innovation and **business growth** [Forbes]

Diverse problem solvers **outperform high ability problem solvers** [Hong & Page]

Companies with diverse executive boards have higher earnings and **returns on equity** [McKinsey]

Multicultural social networks **promote creativity** [Harvard Business School]

Diversity 👍




Information Processing Theory

Mixture of cultural/educational backgrounds
+ access to different networks/broader information
=> creativity, adaptability, & problem solving skills.
[Salancik & Pfeffer]

Today: diversity in open source software (OSS) GitHub teams

Different settings

Geographic & cultural dispersion

Online communities & distributed comm. channels

Different methods

Quantitative; large-scale trace data

Today: gender & tenure diversity in open source software (OSS) GitHub teams

Gender diversity
= mix women/men

simplifying assumption: gender is binary

The "hacker" culture is male-dominated and unfriendly to women [Turkle]

Women are <10% in OSS [Robles et al]

Reports of active discrimination and sexism towards women [Nafus]

Today: gender & tenure diversity in open source software (OSS) GitHub teams

Tenure diversity
= mix junior/senior

The "onion" structure of OSS: small (stable) core + large (loose) periphery [Ducheneaut]

High turnover [Robles & Gonzalez-Barahona]

Today: gender & tenure diversity in open source software (OSS) GitHub teams

World's largest open source community

Trace data available @ghtorrent [Gousios et al]

Today: gender & tenure diversity in open source software (OSS) GitHub teams

Theoretical

Technical

OSS as meritocracy; contribution quality as main driver of impression formation
[Dabbish et al, Marlow et al]

Today: gender & tenure diversity in open source software (OSS) GitHub teams

Theoretical

Technical

Demographics are less salient in OSS [Riordan & Shore]

Today: gender & tenure diversity in open source software (OSS) GitHub teams

Theoretical

Technical

Anyone can contribute to any repository.
Who's on a team?

Today: gender & tenure diversity in open source software (OSS) GitHub teams

Theoretical

Technical

Gender is not explicitly recorded

Today: gender & tenure diversity in open source software (OSS) GitHub teams

Theoretical

Technical

People contribute under multiple aliases

@b_vasilescu @baishakhir @MarkvandenBrand @aserebrenik @devanbu @vlfilkov

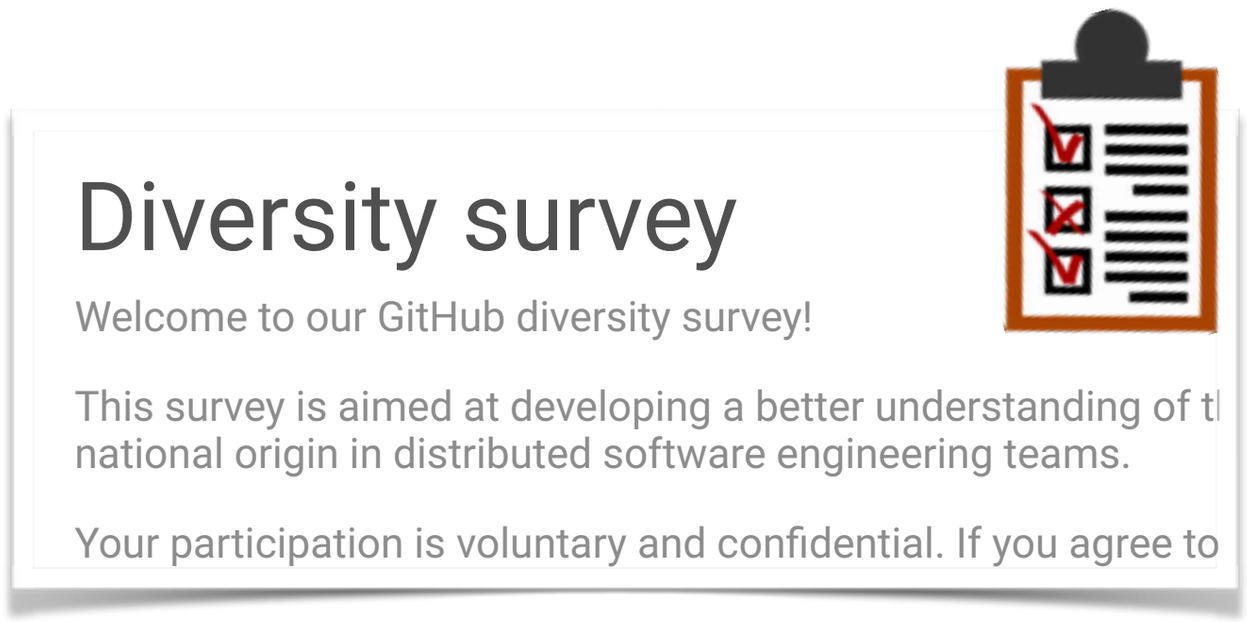
Today: gender & tenure diversity in open source software (OSS) GitHub teams

Theoretical

Technical

How to analyze such large-scale longitudinal trace data?

Approach: mixed methods



+

[Vasilescu et al, CHASE'15]

http://bvasiles.github.io/papers/chase15.pdf

Survey

4,500 invitations, 816 responses

What constitutes a team?

Which differences do people recognize among team members?

Does diversity matter?

[Vasilescu et al, CHASE'15]

http://bvasiles.github.io/papers/chase15.pdf

@b_vasilescu @baishakhir @MarkvandenBrand @aserebrenik @devanbu @vlfilkov

Survey

4,500 invitations, 816 responses

What constitutes a team?

The team is everyone

Which differences do people recognize among team members?

Gender is surprisingly salient

Does diversity matter?

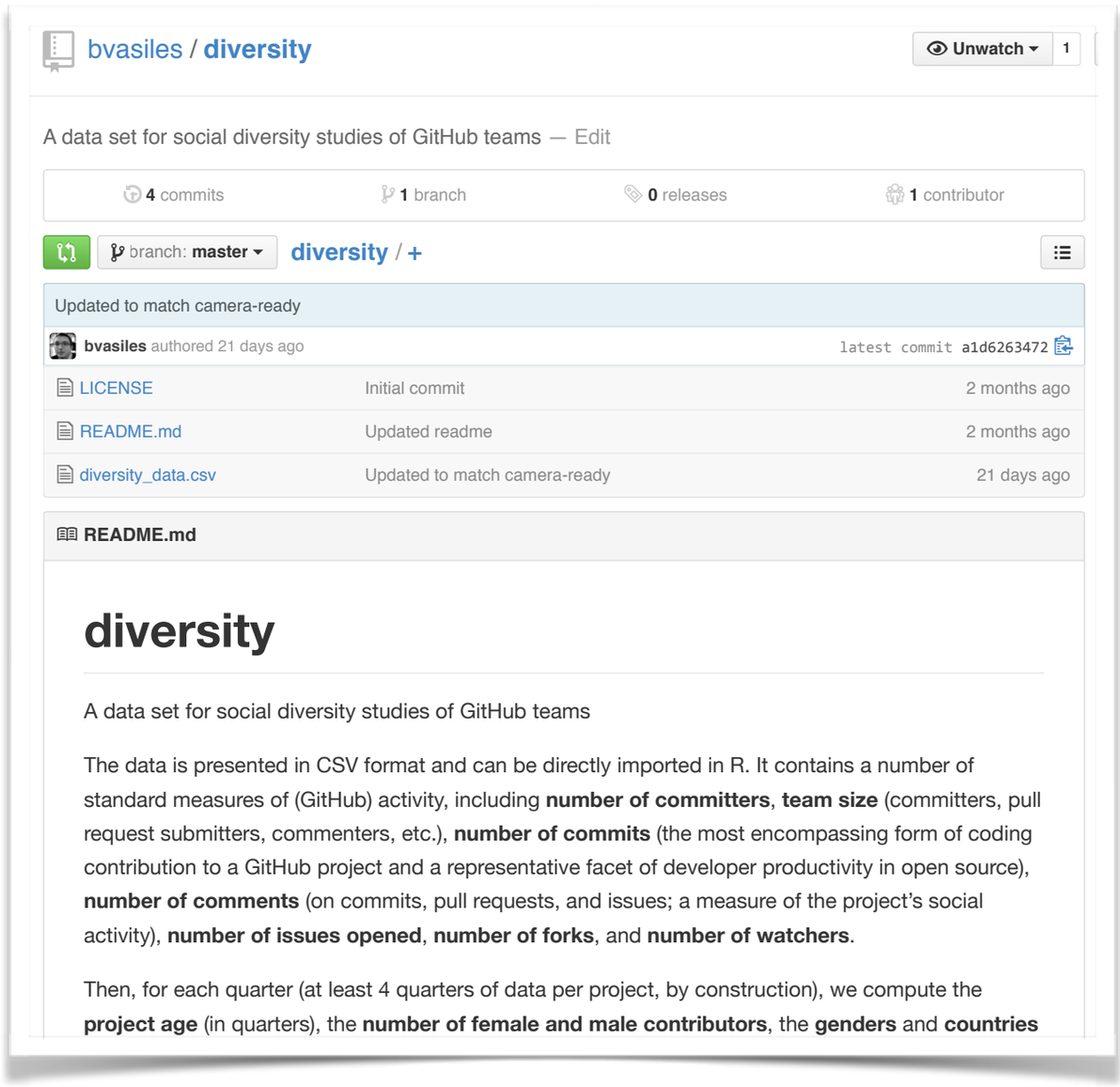
Positive/negative/no effects of diversity

[Vasilescu et al, CHASE'15]

http://bvasiles.github.io/papers/chase15.pdf

Mining

Sample
4K projects



[Vasilescu et al, MSR'15]

- http://bvasiles.github.io/papers/msr_data15.pdf
- https://github.com/bvasiles/diversity

Mining

github SOCIAL CODING

Infer genders (93% precision) [Vasilescu et al, IWC'13]

Sample 4K projects

Bogdan Vasilescu

bvasiles

University of California
Davis, CA
http://bvasiles.github.io
Joined on Jul 3, 2012

Contributions | Repositories | Publi

Popular repositories

- **bvasiles.github.io** — My website
- **diversity** — A data set for social diversity studies of GitHu…
- **flask_assets_tutorial** — Maxime Bouroumeau-Fuseau's tutorial on flas…
- **ghtorrent.org** — The GHtorrent project website
- **ght_unmasking_aliases**

Contributions

Apr May Jun Jul Aug Sep

Bing Maps + Heuristics
http://github.com/tue-mdse/countryNameManager

Bogdan + USA

Name frequency tables for 30 countries
http://github.com/tue-mdse/genderComputer

male

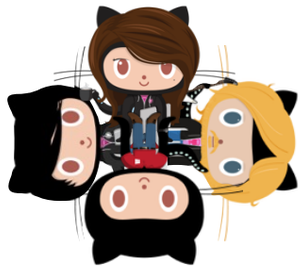
Andrea + Italy = male
Andrea + USA = female

Mining

Sample
4K projects

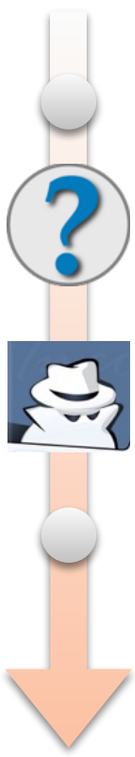
Response

Productivity
(#commits/quarter)



Turnover
(fraction team new
w.r.t. prev. quarter)

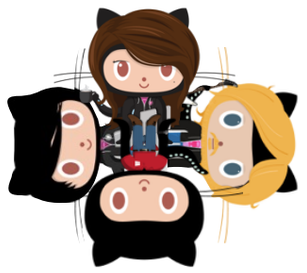
Mining



Sample
4K projects

Response

Productivity
(#commits/quarter)



Turnover
(fraction team new
w.r.t. prev. quarter)

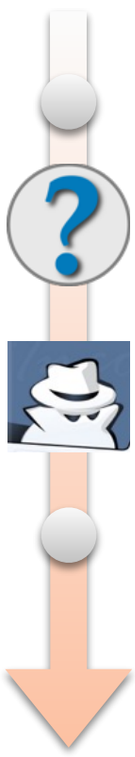
Independent

Gender
diversity
(Blau index)

Tenure diversity
(coeff. variation)

- project
- overall coding

Mining



Sample
4K projects

Response

Productivity
(#commits/quarter)

Turnover
(fraction team new
w.r.t. prev. quarter)

Independent

Gender
diversity
(Blau index)

Tenure diversity
(coeff. variation)

- project
- overall coding

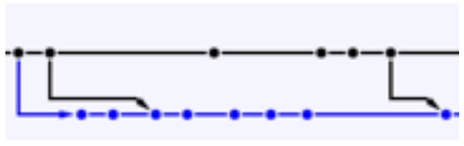
Controls

Team size

Time

Project age

Project activity



...

@b_vasilescu @baishakhir @MarkvandenBrand @aserebrenik @devanbu @vlfilkov

Analysis

Nesting: projects

Project	Created on	Project age	Total #commits	#Forks	Time	#Commits	#Comments	Team size	Gender diversity	Commit tenure diversity	Turnover
A	2011-02-15	12	557	51	Q2	47	26	9	0.25	0.47	0.67
					Q5	19	12	10	0.00	0.93	0.75
					Q6	7	13	12	0.25	0.54	0.67
					Q7	56	53	20	0.00	0.56	0.87
					…						
B	2010-09-21	11	2075	578	Q4	71	169	83	0.03	0.66	0.87
					Q5	116	219	93	0.05	0.73	0.56
					Q6	186	367	119	0.06	0.80	0.86
					Q7	129	453	114	0.08	0.85	0.82
					…						

Analysis

Nesting: projects
Cross-classification: quarters

Project	Created on	Project age	Total #commits	#Forks	Time	#Commits	#Comments	Team size	Gender diversity	Commit tenure diversity	Turnover
A	2011-02-15	12	557	51	Q2	47	26	9	0.25	0.47	0.67
					Q5	19	12	10	0.00	0.93	0.75
					Q6	7	13	12	0.25	0.54	0.67
					Q7	56	53	20	0.00	0.56	0.87
					…						
B	2010-09-21	11	2075	578	Q4	71	169	83	0.03	0.66	0.87
					Q5	116	219	93	0.05	0.73	0.56
					Q6	186	367	119	0.06	0.80	0.86
					Q7	129	453	114	0.08	0.85	0.82
					…						

Analysis

Nesting: projects
Cross-classification: quarters

Linear mixed-effects (hierarchical) models

Project	Created on	Project age	Total #commits	#Forks	Time	#Commits	#Comments	Team size	Gender diversity	Commit tenure diversity	Turnover
A	2011-02-15	12	557	51	Q2	47	26	9	0.25	0.47	0.67
					Q5	19	12	10	0.00	0.93	0.75
					Q6	7	13	12	0.25	0.54	0.67
					Q7	56	53	20	0.00	0.56	0.87
					…						
B	2010-09-21	11	2075	578	Q4	71	169	83	0.03	0.66	0.87
					Q5	116	219	93	0.05	0.73	0.56
					Q6	186	367	119	0.06	0.80	0.86
					Q7	129	453	114	0.08	0.85	0.82
					…						

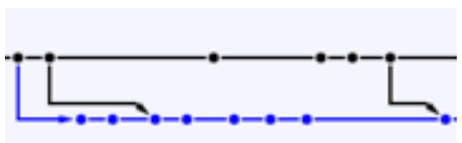
Results

Productivity
(#commits/quarter)

\+ \+ \- \-

Gender diversity

Team size



Overall project activity

Forks

Project age

Tenure diversity

…

@b_vasilescu @baishakhir @MarkvandenBrand @aserebrenik @devanbu @vlfilkov

Results

Productivity
(#commits/quarter)

\+ all team sizes

Gender diversity

\+ mid-size & large teams

Tenure diversity

\+ Team size

\+ Overall project activity

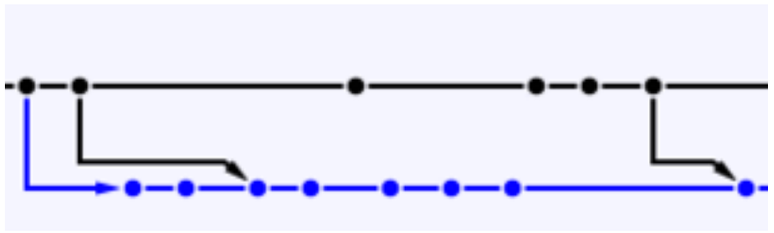
\- Forks

\- Project age

…

Results

Productivity
(#commits/quarter)



+

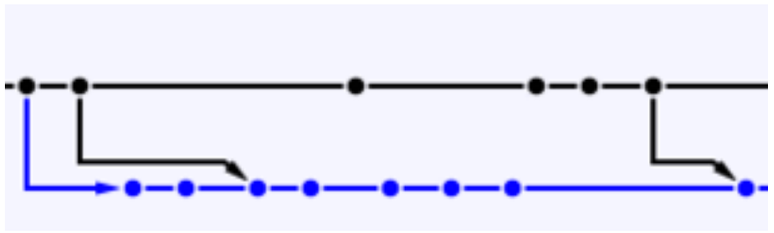
Gender diversity

+

Tenure diversity

Results

Productivity
(#commits/quarter)



Gender diversity

Turnover (fraction team new w.r.t. prev. quarter)

Tenure diversity

+

+

+

The takeaway

Today: diversity in open source software (OSS) GitHub teams

Different settings

Geographic & cultural dispersion

Online communities & distributed comm. channels

Different methods

Quantitative; large-scale trace data

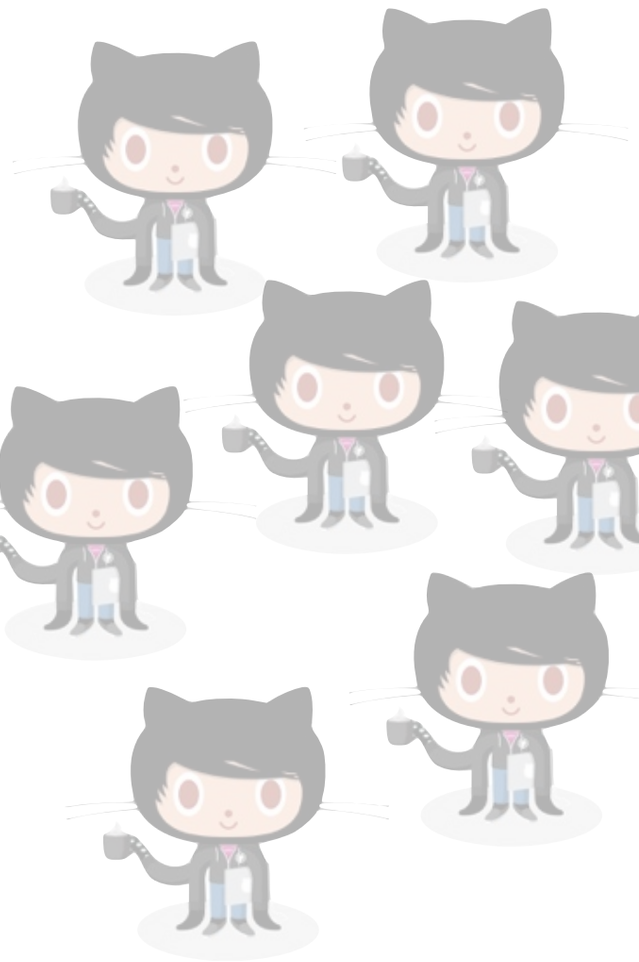
The takeaway

Toc

D

Geograph
cultural disp

Which is more effective?



@b_vasilescu @baishakhir @MarkvandenBrand @aserebrenik @devanbu @vlfilkov

The takeaway

Today: diversity in open source software (OSS) GitHub teams

Different settings

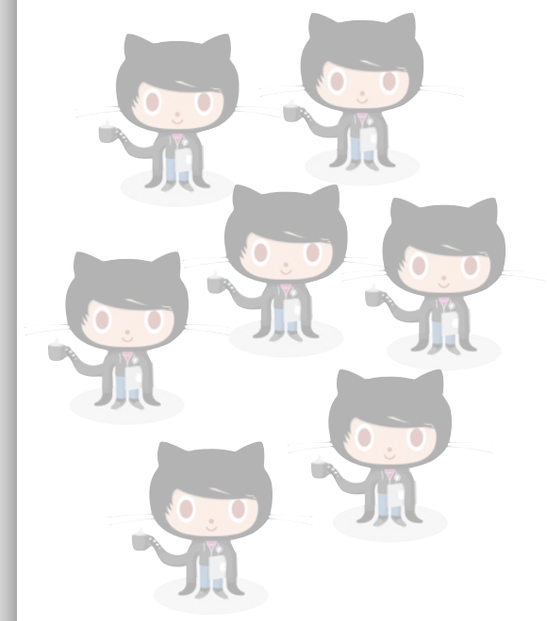
Geographic & cultural dispersion

Online communities & distributed comm. channels

Different methods

Quantitative; large-scale trace data

Which is more effective?



Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark van den Brand, Alexander Serebrenik, Prem Devanbu, Vladimir Filkov

NWO Netherlands Organisation for Scientific Research

NSF

AFOSR AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

@b_vasilescu @baishakhir @MarkvandenBrand @aserebrenik @devanbu @vlfilkov

Octocats from the GitHub Octodex https://octodex.github.com/