

# Developer Onboarding in GitHub: Effects of Social Links & Language Experience

Casey Casalnuovo, Bogdan Vasilescu,  
Prem Devanbu, Vladimir Filkov

Why then the world's mine oyster,  
Which I with sword will open.

W. Shakespeare



In GitHub Many Oysters (Projects)  
Lie Waiting to Be Opened

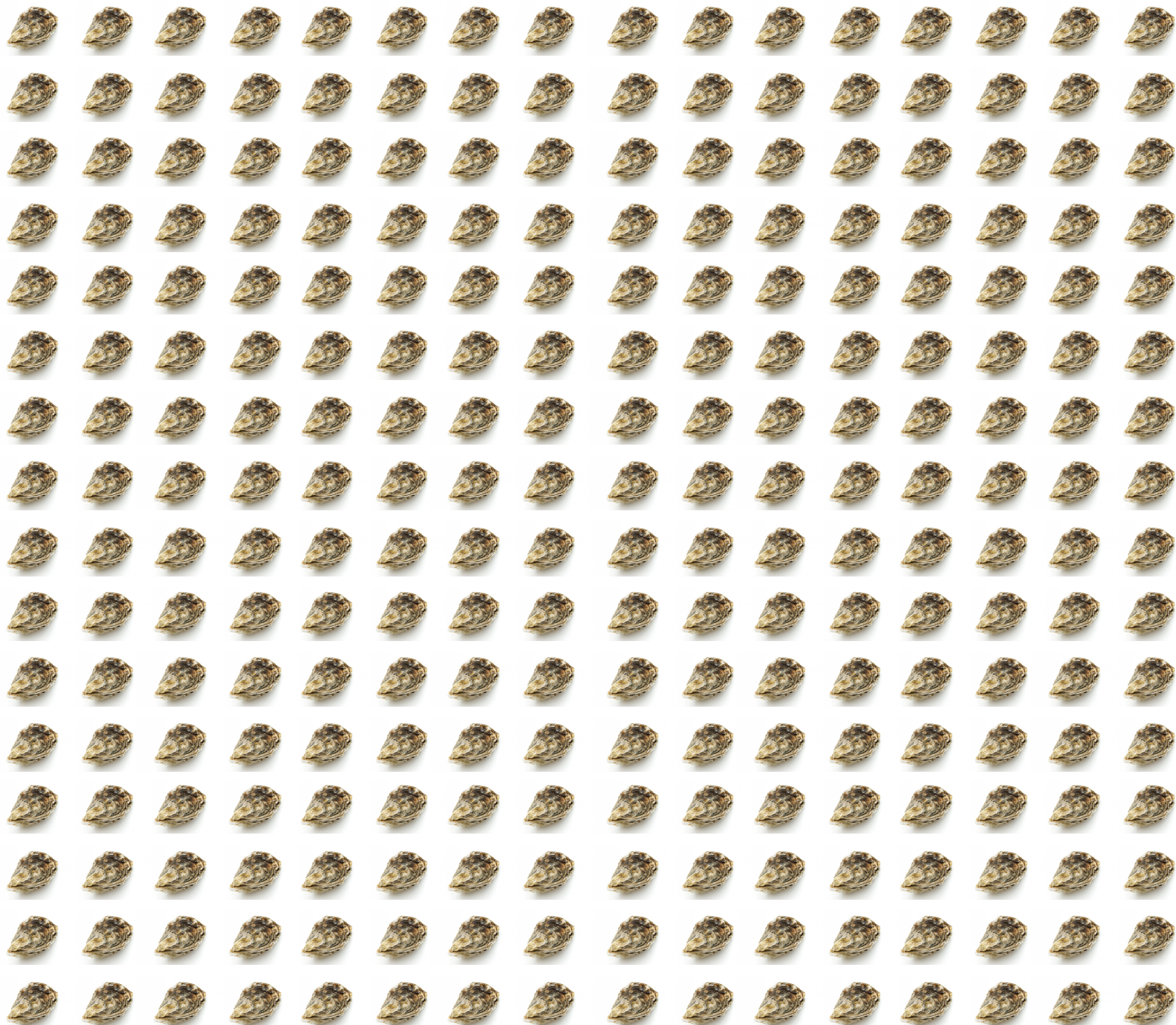


# What Opportunities Await GitHub Coders?

- Fun
- Knowledge
- Employment
- Fame
- Fortune

# Great, I know How to Code

- Now, show me the oysters...



# Shoot, too many!

- How to sort through them?





# Which projects to join?



- Popularity
- Social connections
- Technical familiarity





# Social





# Social

Started in:

 = 2010

 = 2011

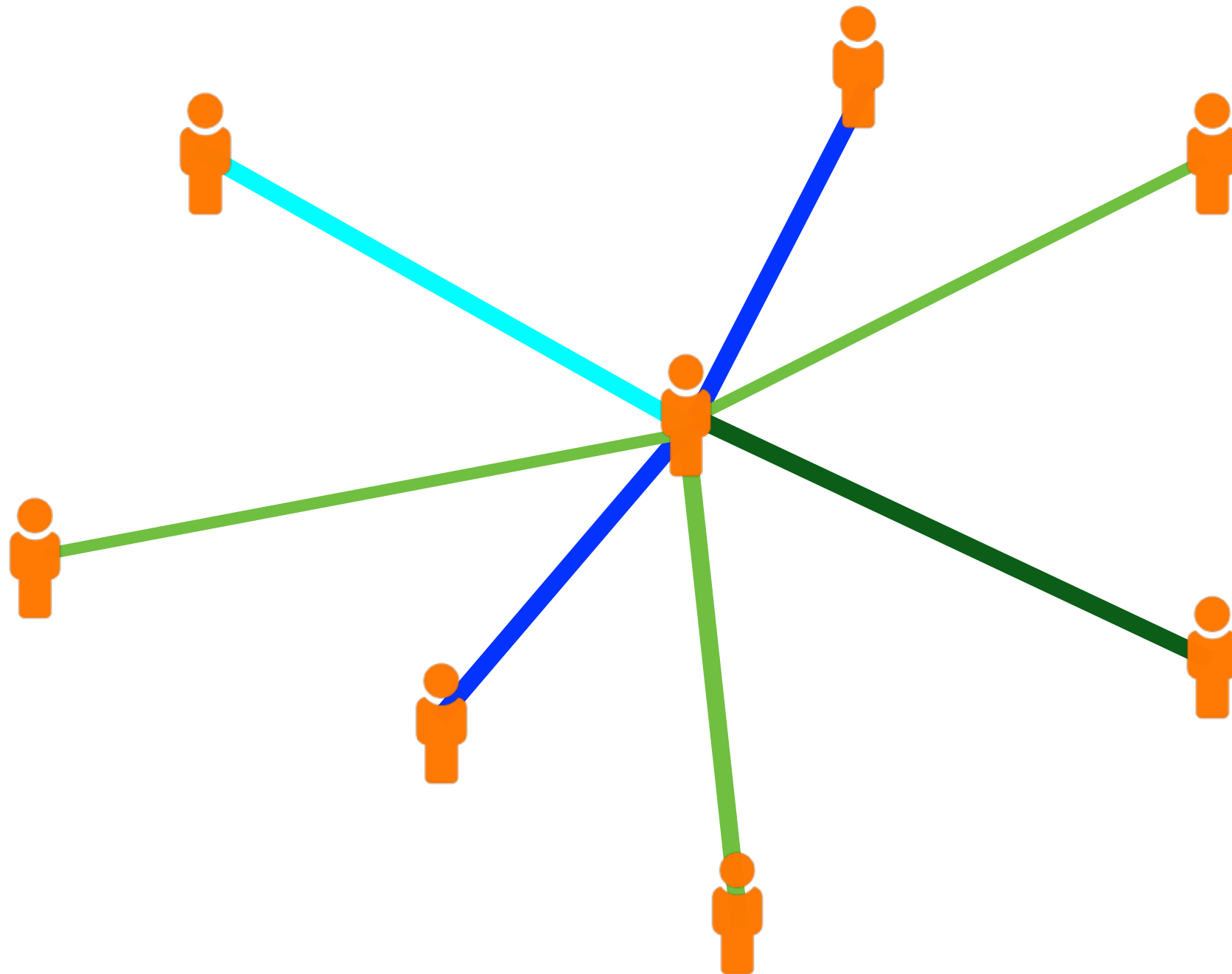
 = 2012

 = 2013

Shared Projects:

 = 2

 = 3



# Technical


A word cloud featuring various programming languages and technologies. The words are arranged diagonally from the top-left to the bottom-right. The largest word is 'JavaScript' in red. Other prominent words include 'Python' in brown, 'Ruby' in green, 'Java' in blue, 'PHP' in light green, 'CSS' in lime green, 'C++' in green, and 'ObjectiveC' in yellow. Smaller words include 'Shell' in orange, 'Scala' in red, 'R' in brown, 'Lua' in purple, 'Clojure' in purple, 'Perl' in red, 'CoffeeScript' in red, 'VimL' in purple, and 'TeX' in purple.


Python Ruby CSS JavaScript PHP C++ ObjectiveC  
Shell Scala R Lua Clojure  
CoffeeScript Perl VimL TeX Java

How can we quantify these social and technical effects during onboarding in GitHub projects?

# Research Questions

 Do developers select projects with past social links preferentially?

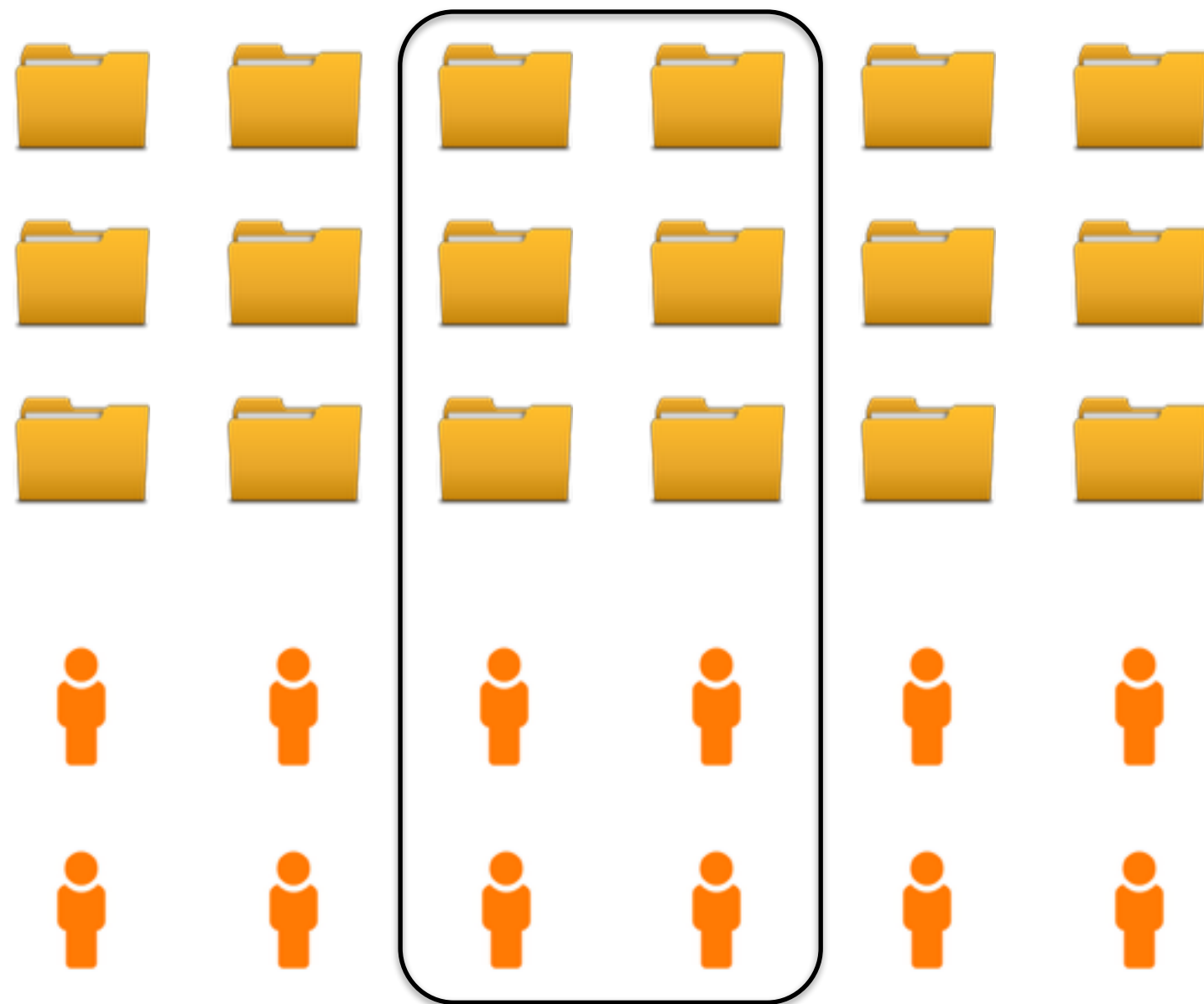
 How does language experience and strength of social connection affect productivity in the initial joining period?

 How does language experience and strength of social connection affect productivity in the long term?

# Methodology

- User Selection + Project Selection from GHTorrent
- De-Aliasing
- Prior Experience with Project Languages
- Social Links Metric
- Combinatorial and Statistical Modeling

# User and Project Selection





# User and Project Selection

- From GHTorrent Selected Prolific Devs:  
500+ commits, 5 years on GitHub, at least 10 projects

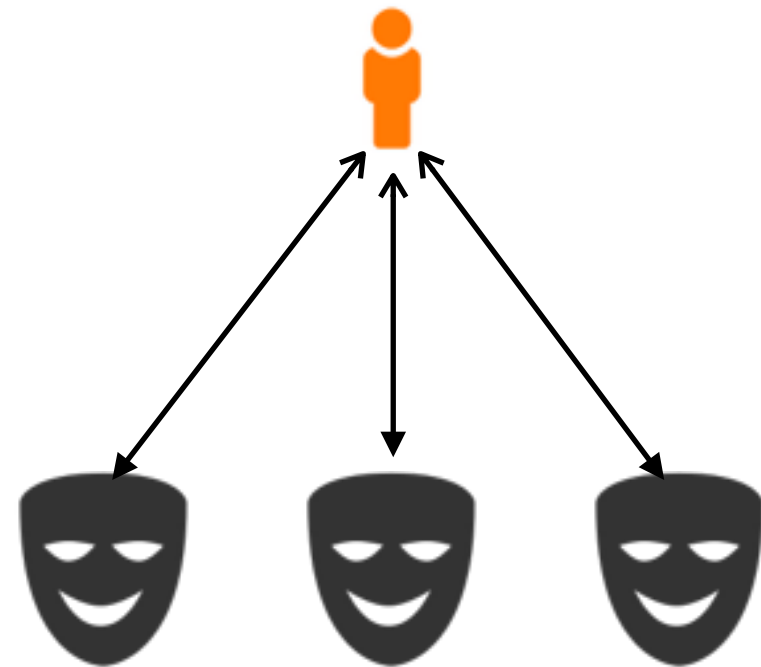
Description	GHTorrent	404 Not Found and Log Errors
# Projects	65.280	58.092
# Prolific Developers	1.274	1.255

- Cloned and parsed the git logs of all their repositories not marked as forks.

# Aliasing Problem

- One developer may use different emails and user names.
- To more accurately identify people and not names, we combine username - email pairs to a single person id.

Person ID = 29



marat yakupov  
marat yakupov  
moadib

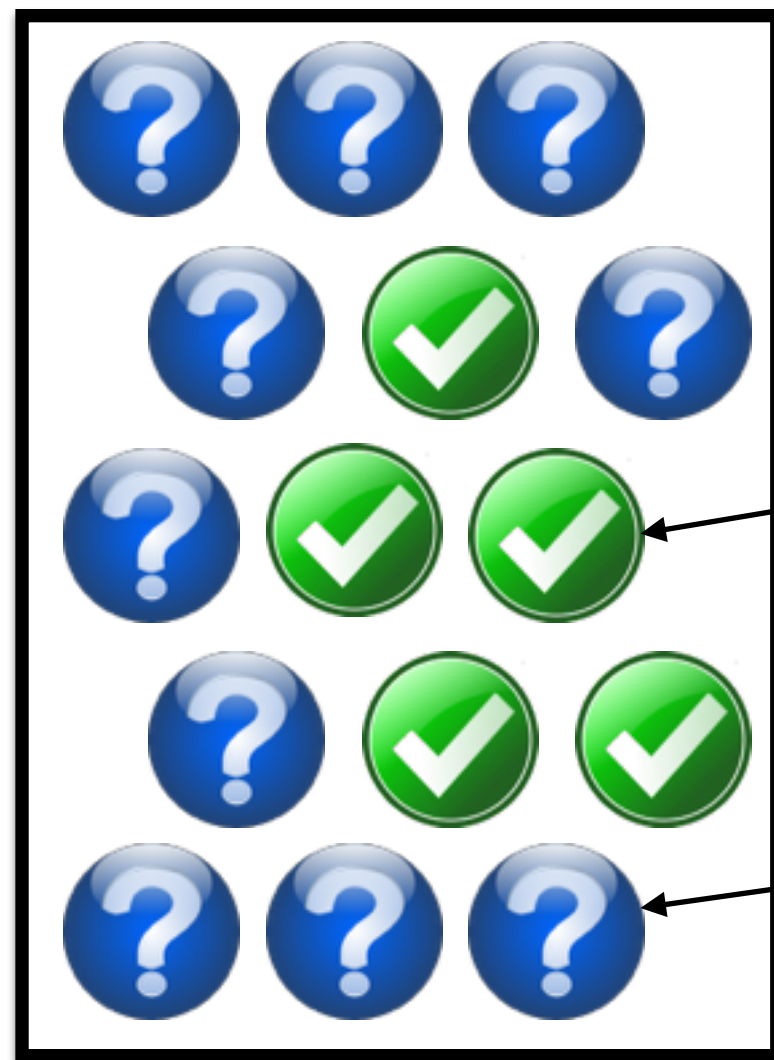
moadib73rus@gmail.com  
markosstudio@gmail.com  
moadib73rus@gmail.com

# RQ1: Do Developers preferentially join projects with prior social connections?

- A developer looking at the pool of available projects to join, finds that some contain prior social connections (i.e., people that they have already been around in other projects).
- Do developers join these projects **more frequently than expected by chance?**

# Hypergeometric Test

~1/3 Have links



GitHub from a  
Developer's  
Perspective

Projects With Social Links

Projects With No Links

Random  
Sample



Expect:  
 $\frac{1}{3}$  Have  
links

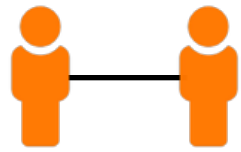
Developer's  
Actual Choice



Get more  
than 1/3?



Reject Random if  $p < 0.05$





# RQ1: Do Developers prefer joining projects where there are social connections?

Description	Reject random	Not able to reject random	Percentage
# Developers	1081	119	90,1%
# Joining Events	4199	2854	59,5%

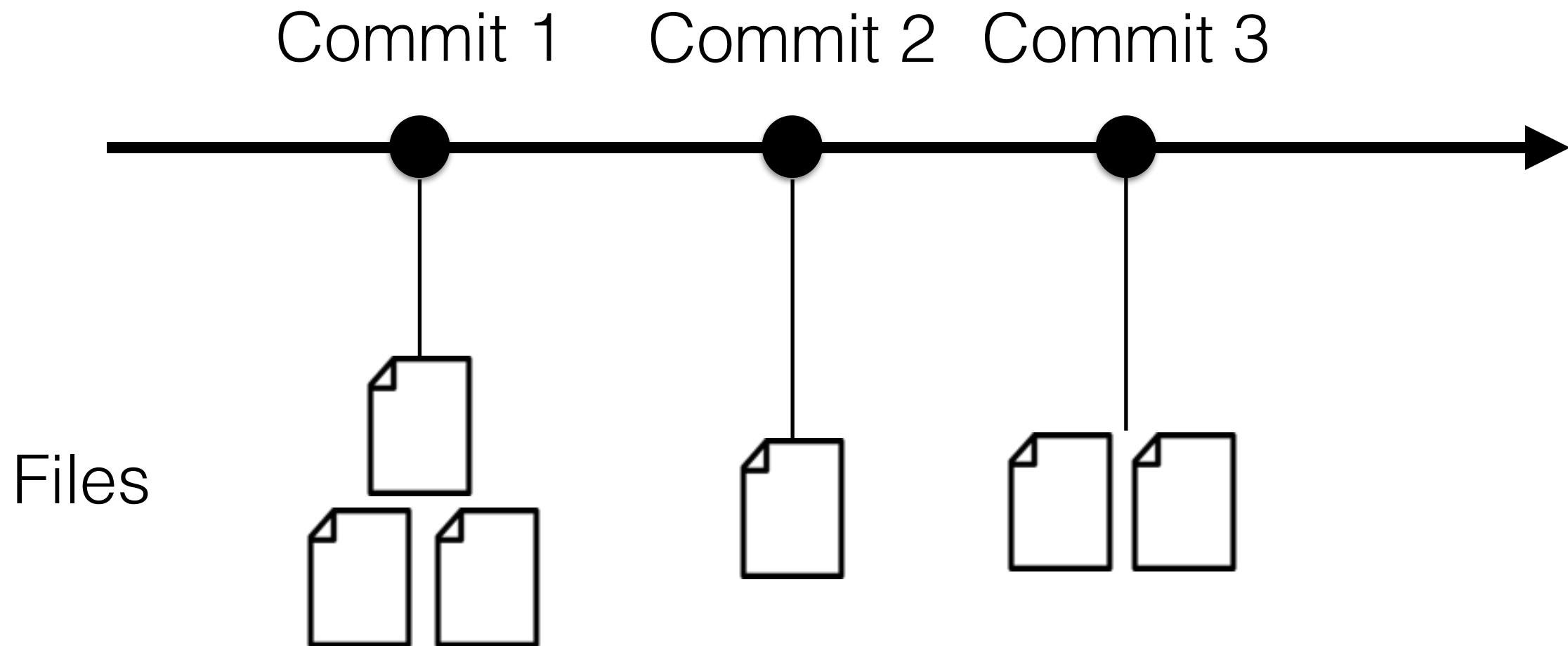


# RQ2 and RQ3:

## Productivity=f(Experience,Links)

- Response: Productivity  or 
- Independent Variables:
  - Language Experience, Strength of Social Connection to Project.
- Controls: Founder, Time Period, #Other projects, total productivity

# Productivity



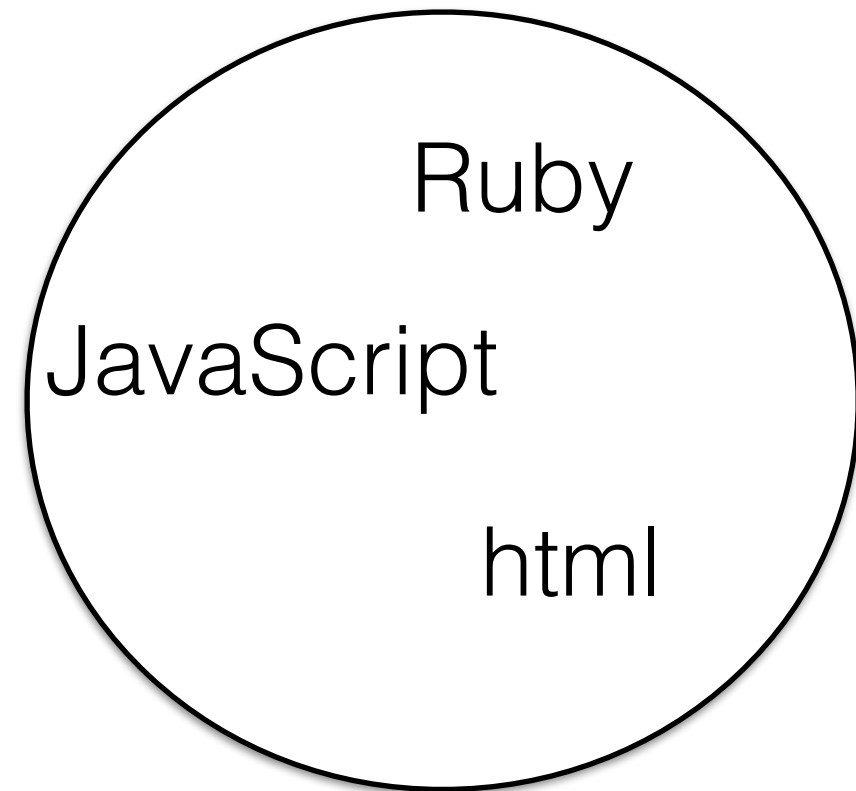
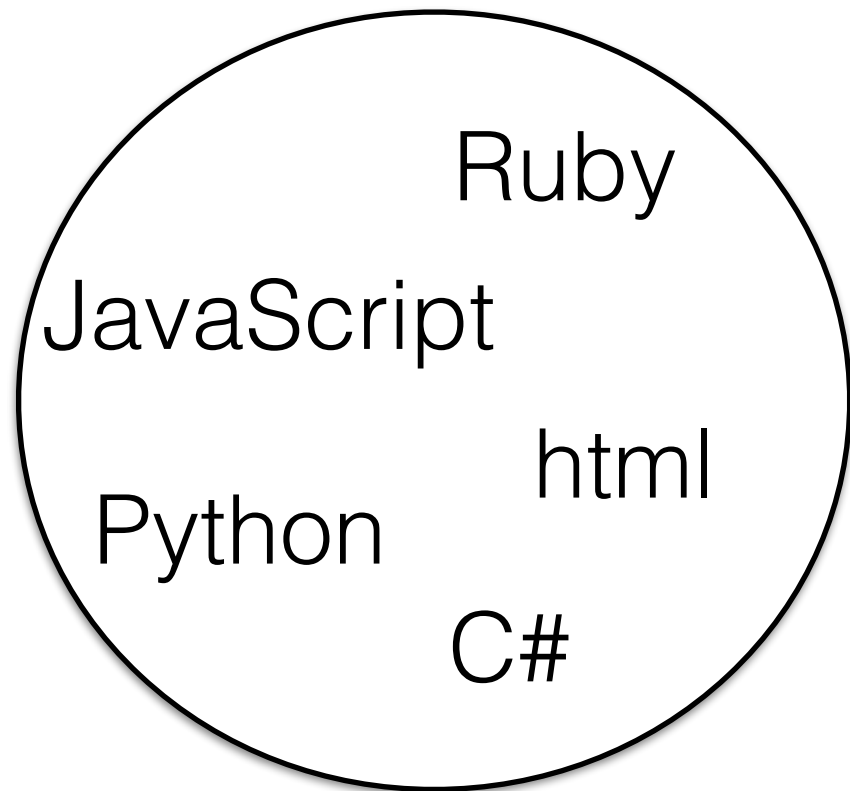
Too coarse a granularity at the commit level.

Lines added and deleted: very noisy.

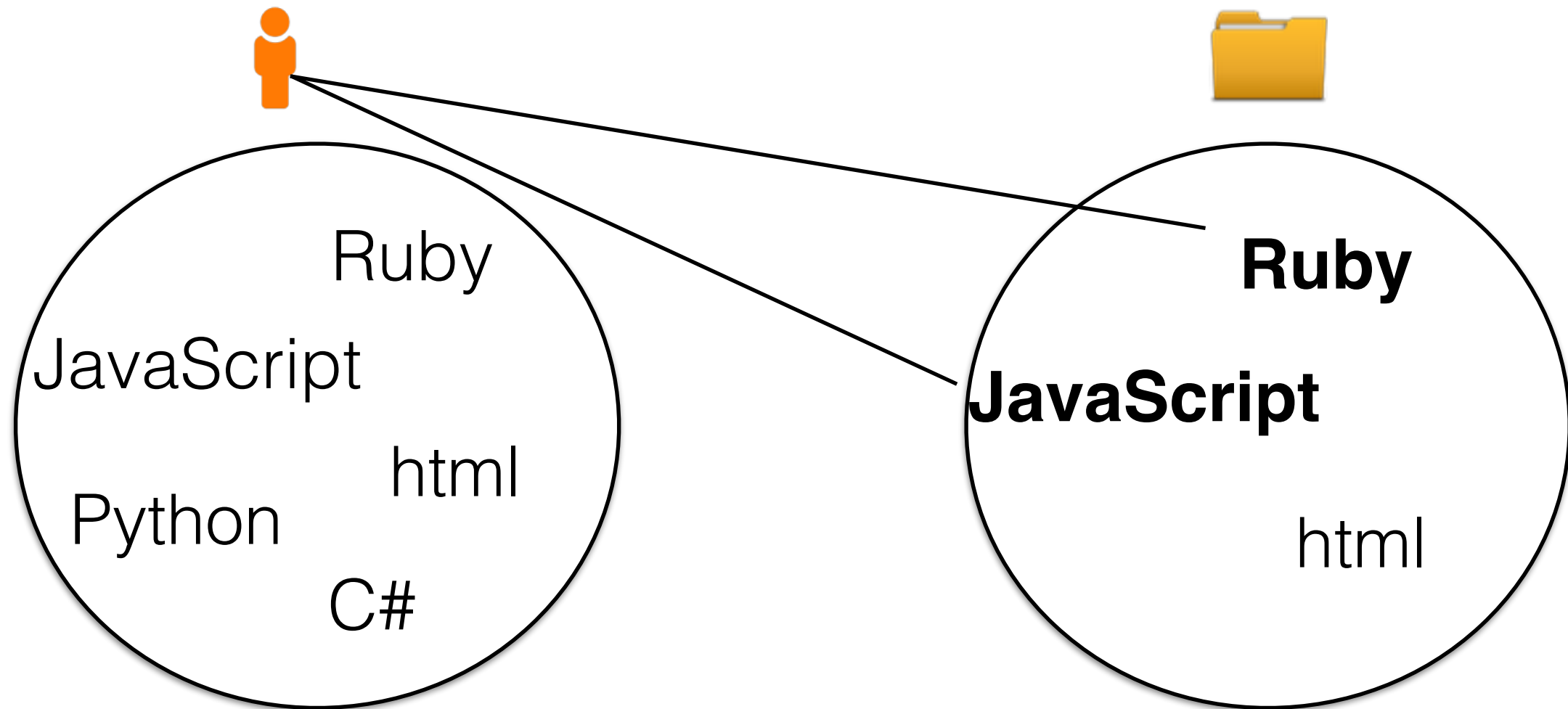
# Prior Language Experience

- Looked at 32 popular languages.
- Language of a file is determined by its extension, and if extension is ambiguous, by context of other files in the project and the project's language tag.

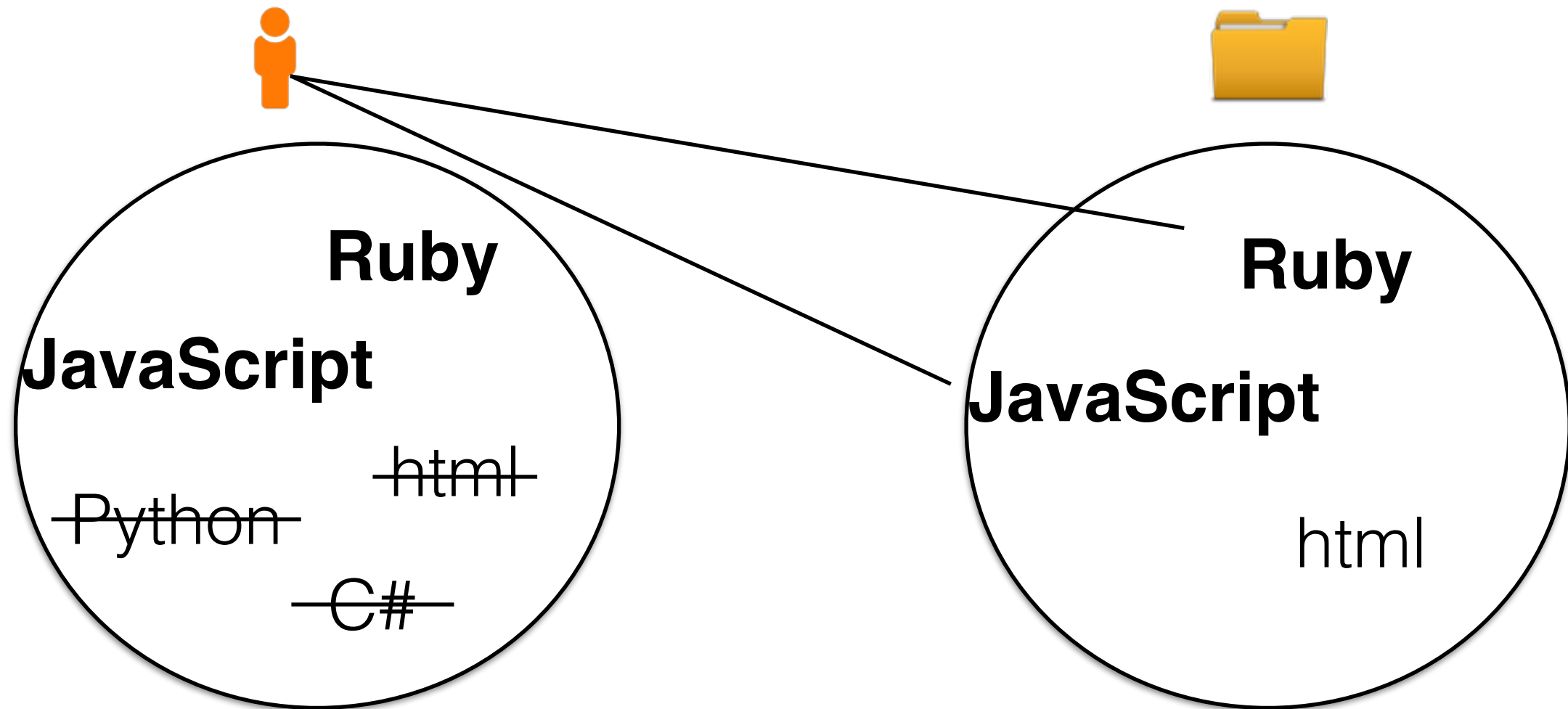
# Language Experience



# Language Experience

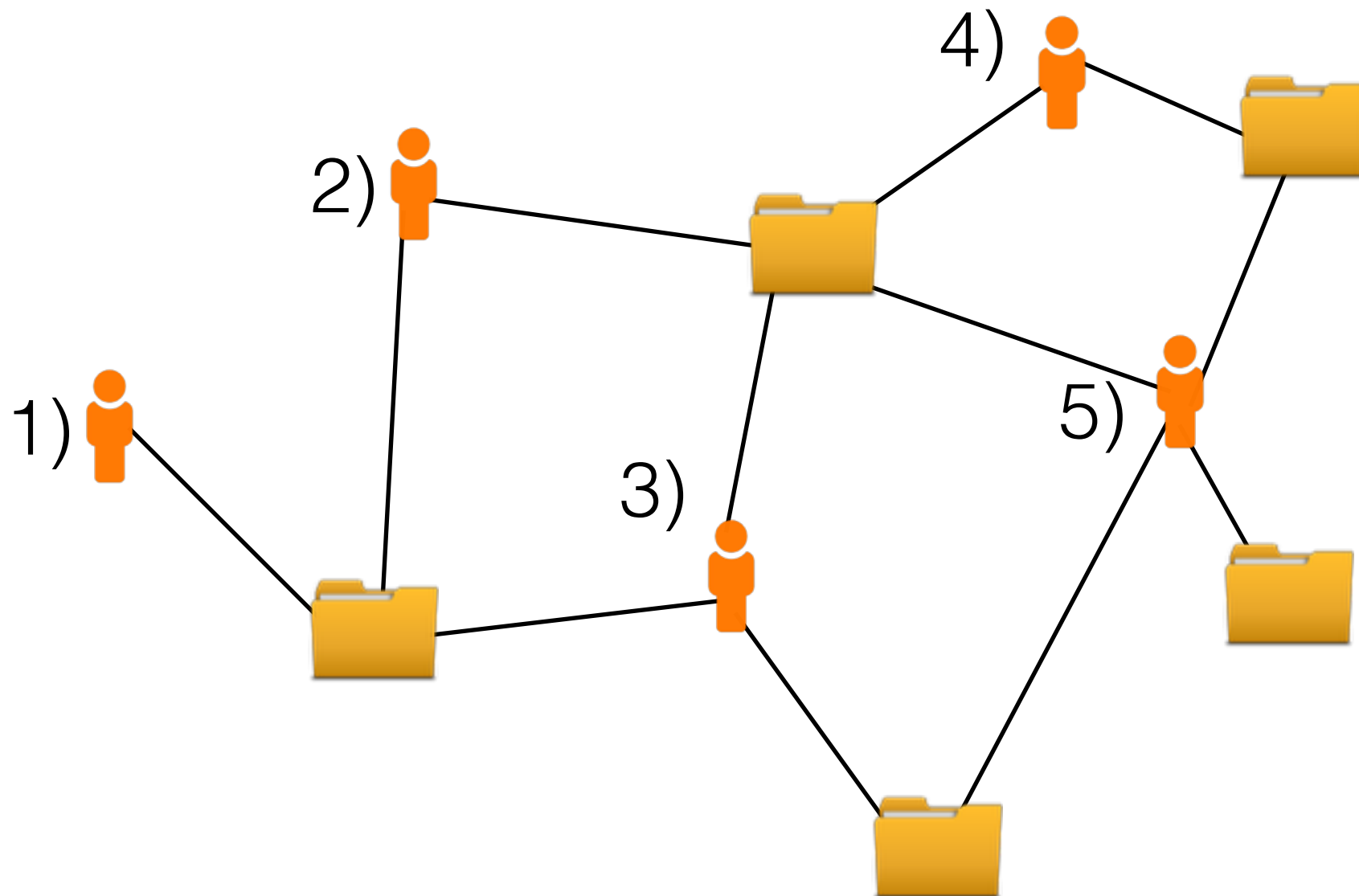


# Language Experience



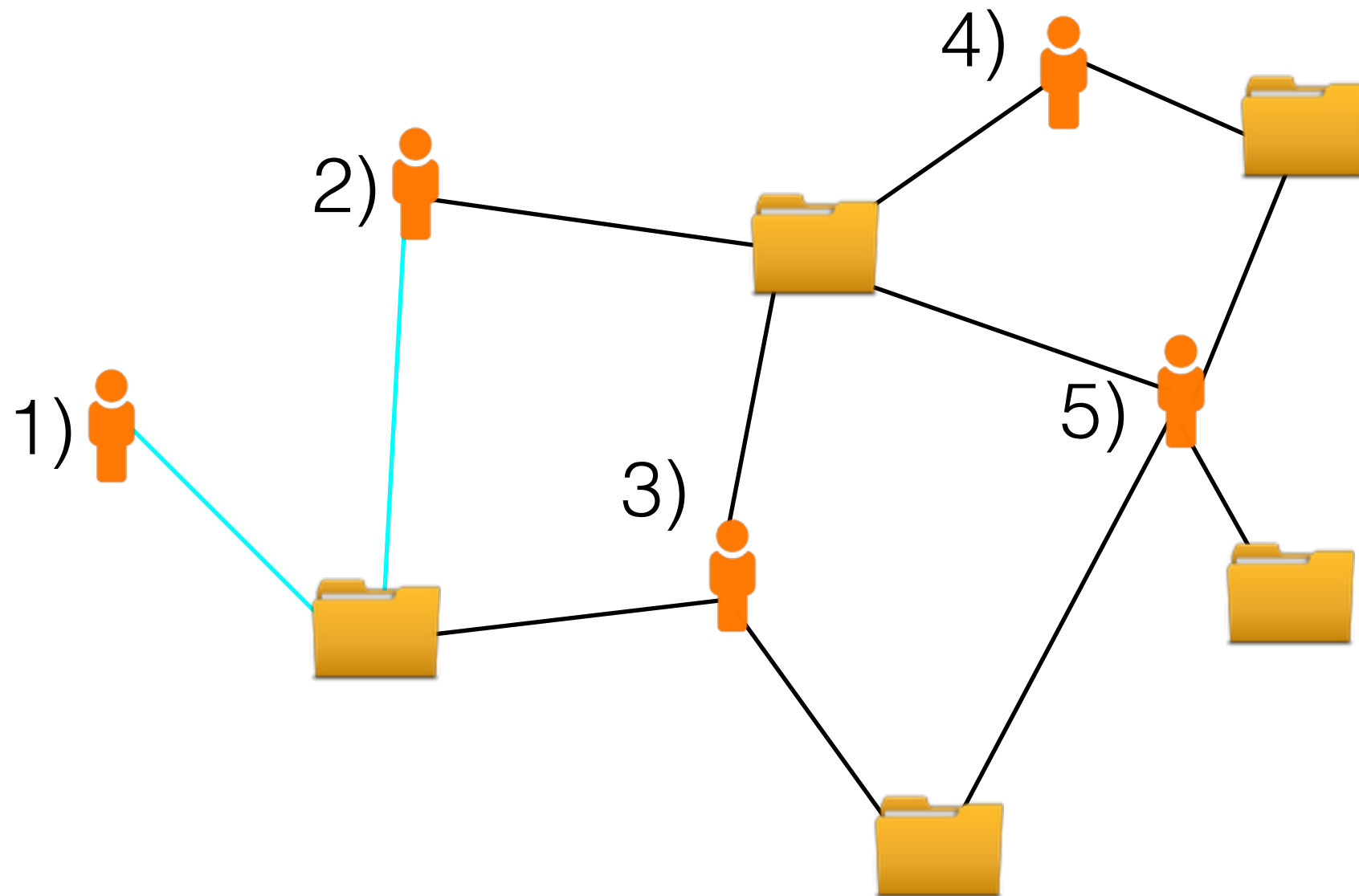
# Prior Social Links

Start from bipartite contribution network of developers and projects on Github

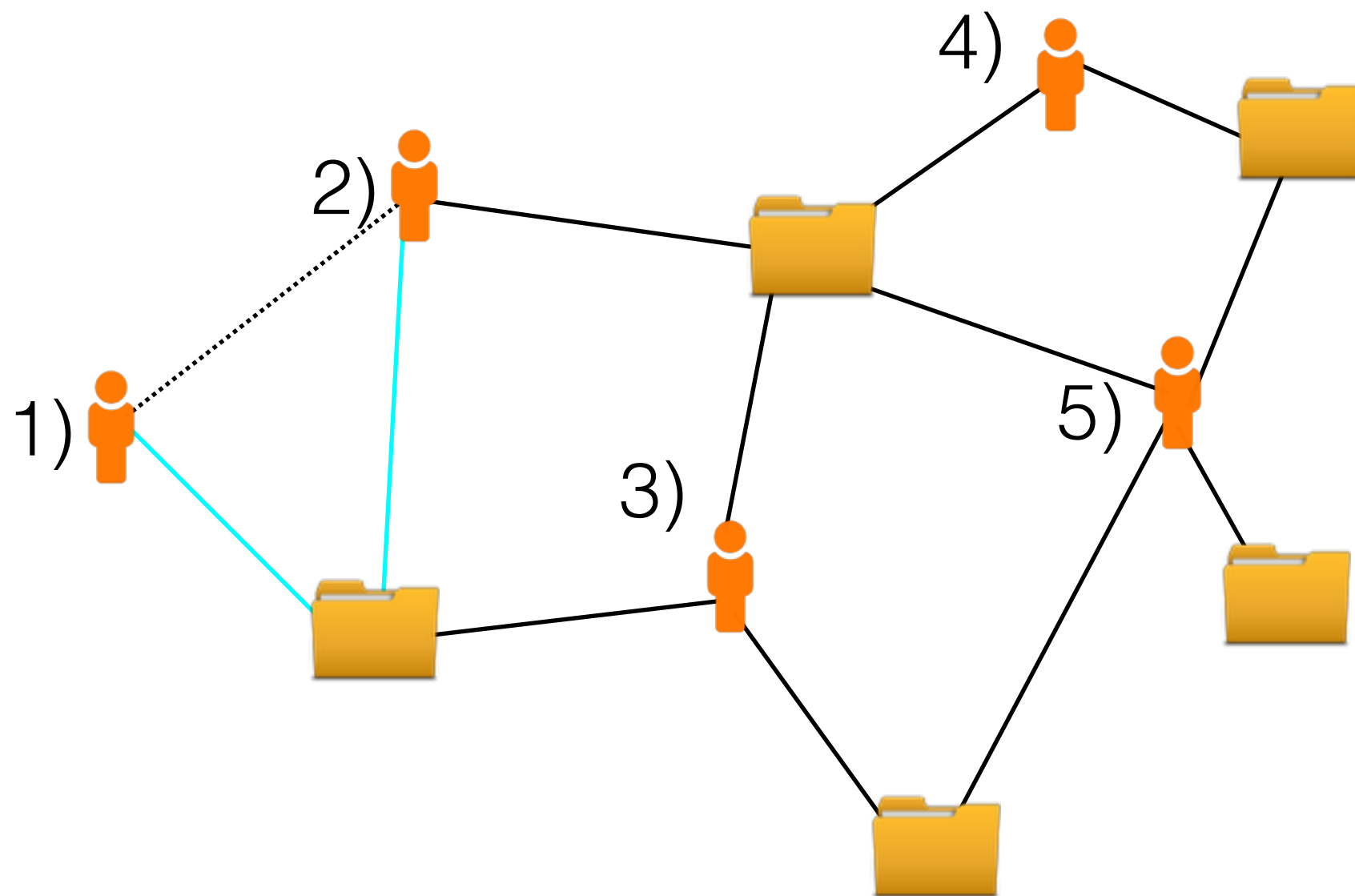




# Contribution Network

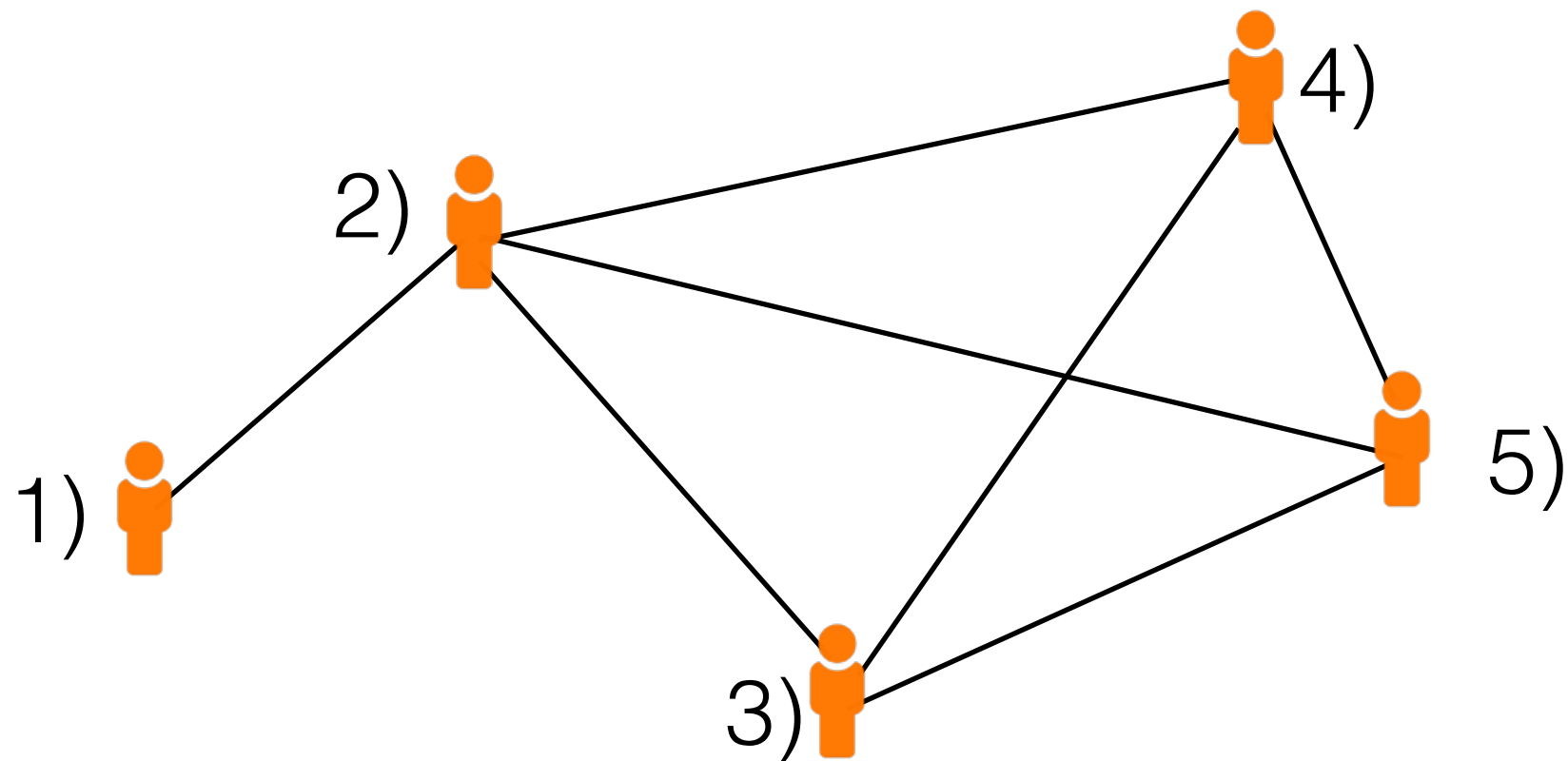


# Contribution Network



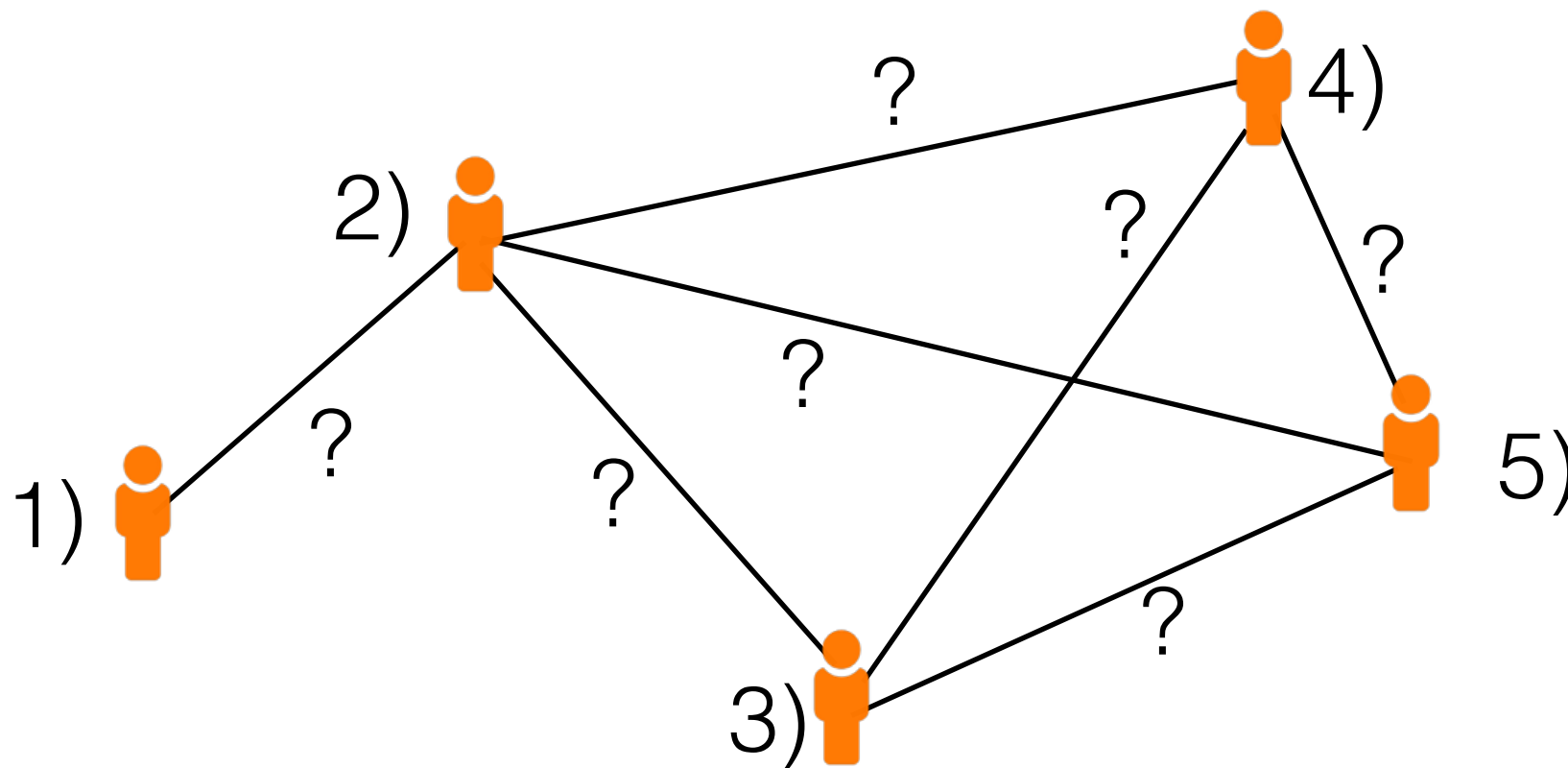
# Contribution Network to Social Network

Can answer: **Is there a connection?**



# Contribution Network to Social Network

Next: **How Strong is the connection?**



# Social Link Strength

- Factors that effect the strength of connection between 2 developers:
  - How many projects do they share?
  - How many people worked in those projects?
  - This may change over time as more projects shared.

# Prior Social Connection

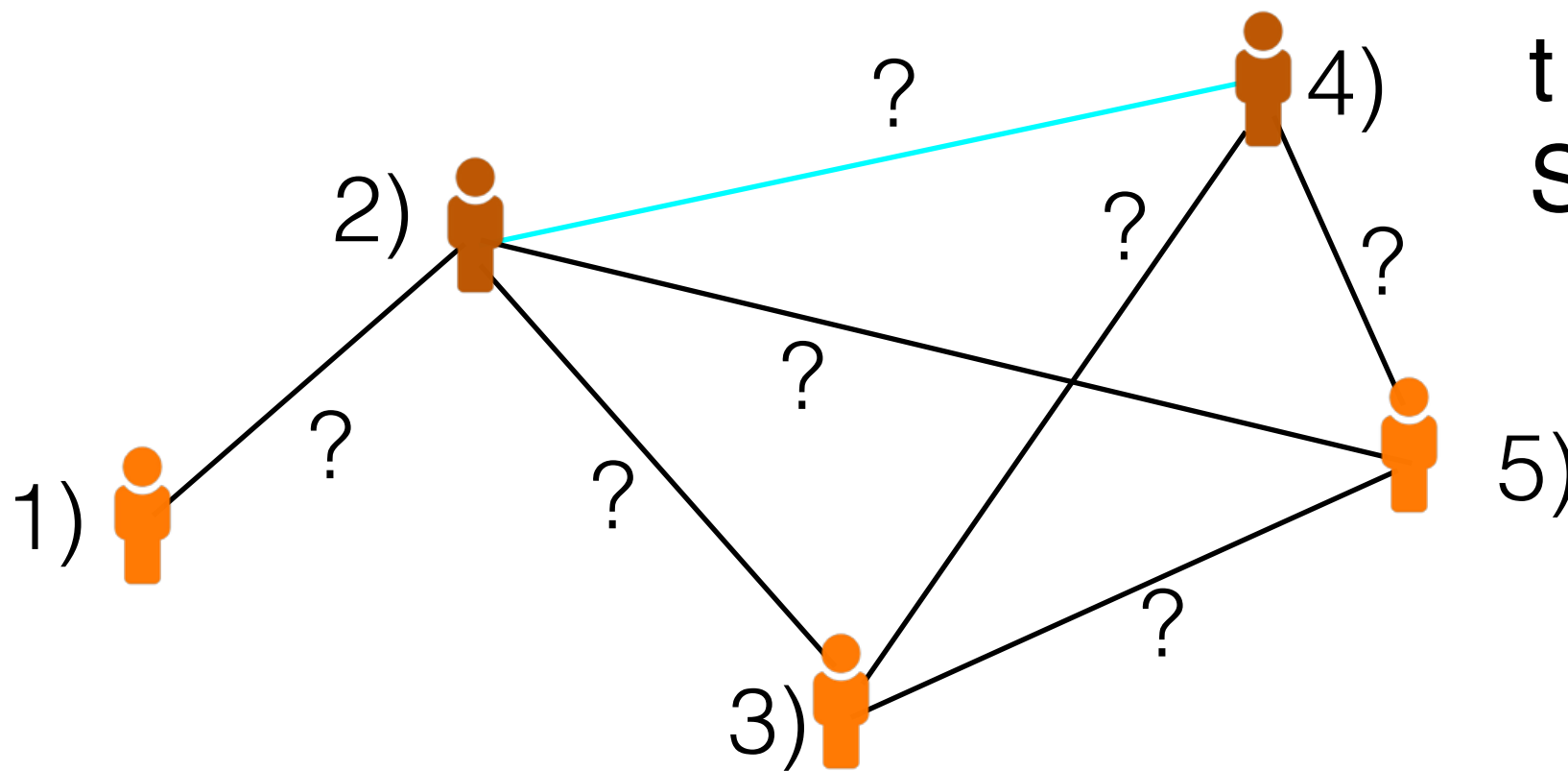
How Strong is the connection?

$$? = \sum_{i=1}^P \frac{1}{S_{i,t}}$$

P = prior shared projects

t = time period

S = Team size of project



Prior connection to a project is the sum of these weights for each existing contributor.



# RQ2: What are the socio-technical effects on initial productivity?



---

Negative Binomial Model

\* =  $p < 0.1$

\*\* =  $p < 0.05$

\*\*\* =  $p < 0.01$

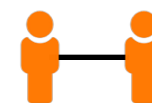
Experience



Is Founder



Has Links



Link Strength







# RQ2: What are the socio-technical effects on initial productivity?

\*\*\*  
↑ 157.3%



---

Negative Binomial Model

\* =  $p < 0.1$

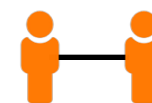
\*\* =  $p < 0.05$

\*\*\* =  $p < 0.01$

Experience



Has Links



Is Founder




Link Strength

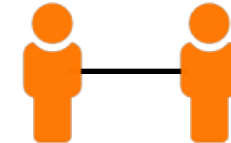




# RQ2: What are the socio-technical effects on initial productivity?

\*\*\*  
 ↑ 157.3%



\*\*\*  
 ↑ 6.2%

---

Negative Binomial Model

\* =  $p < 0.1$

\*\* =  $p < 0.05$

\*\*\* =  $p < 0.01$

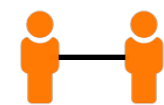
Experience



Is Founder



Has Links




Link Strength

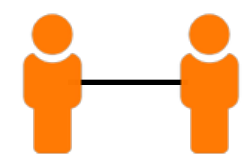






# RQ2: What are the socio-technical effects on initial productivity?

\*\*\*  
 ↑ 157.3%



\*\*\*  
 ↑ 6.2%

\*\*\*  
 ?  ↓ -2%

---

Negative Binomial Model

\* =  $p < 0.1$

\*\* =  $p < 0.05$

\*\*\* =  $p < 0.01$

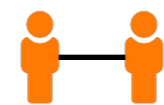
Experience



Is Founder



Has Links

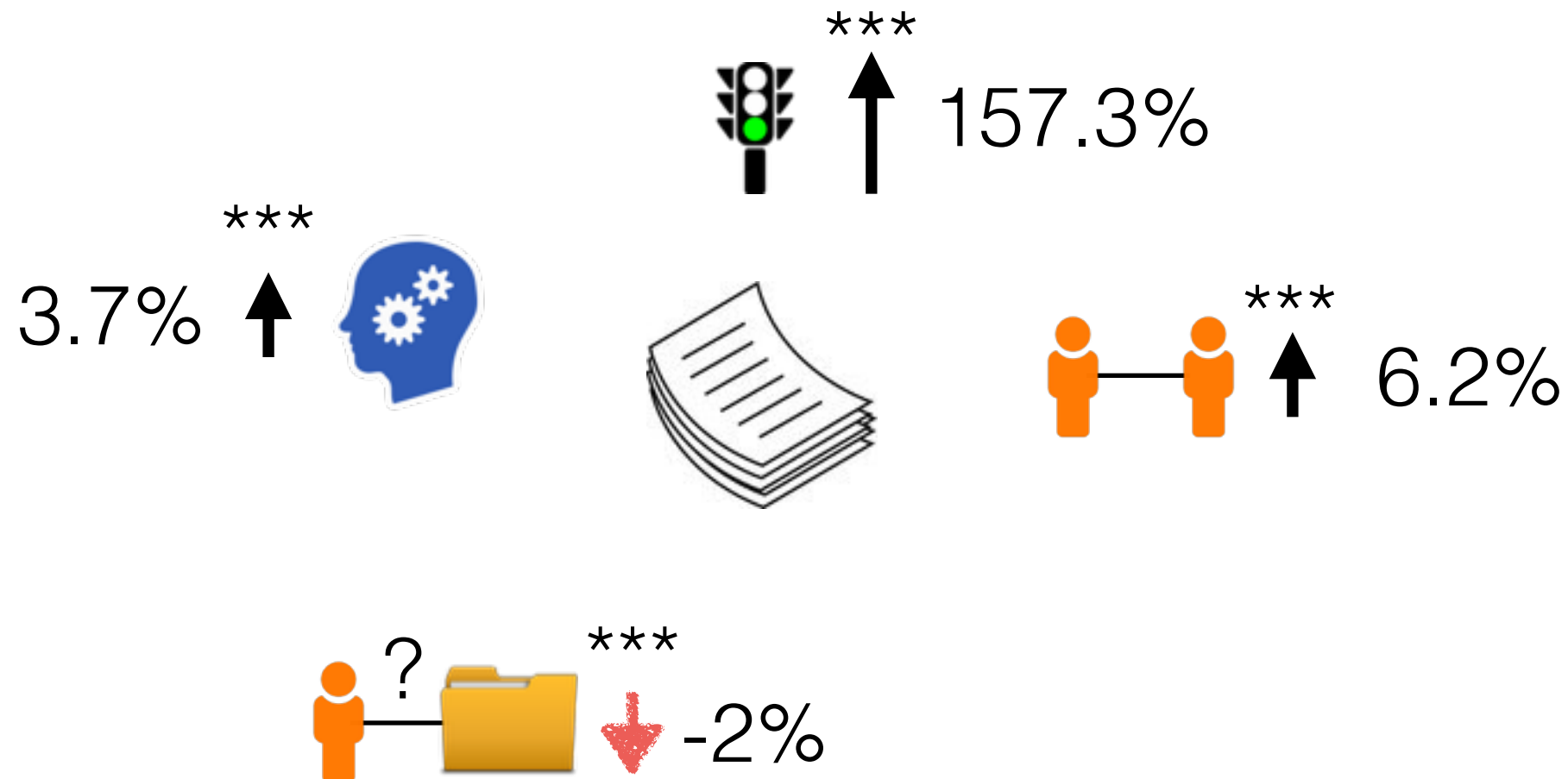


Link Strength





# RQ2: What are the socio-technical effects on initial productivity?



---

Negative Binomial Model

\* =  $p < 0.1$

\*\* =  $p < 0.05$

\*\*\* =  $p < 0.01$

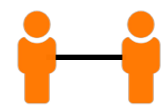
Experience



Is Founder



Has Links



Link Strength



# Initial Productivity

- Both prior language experience and having some link to the project lead to an increase in productivity.
- However, a stronger social link to a project has a small cost to initial productivity.



# RQ3: What are the socio-technical effects on cumulative productivity?



---

Negative Binomial Model

\* =  $p < 0.1$

\*\* =  $p < 0.05$

\*\*\* =  $p < 0.01$

Experience



Has Links



Is Founder



Link Strength

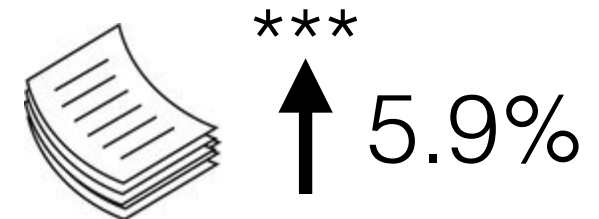
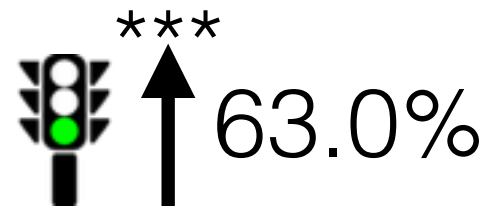


Time period  
joined  
initial file  
changes





# RQ3: What are the socio-technical effects on cumulative productivity?



---

Negative Binomial Model

\* =  $p < 0.1$

\*\* =  $p < 0.05$

\*\*\* =  $p < 0.01$

Experience



Has Links



Time period  
joined



Is Founder



Link Strength

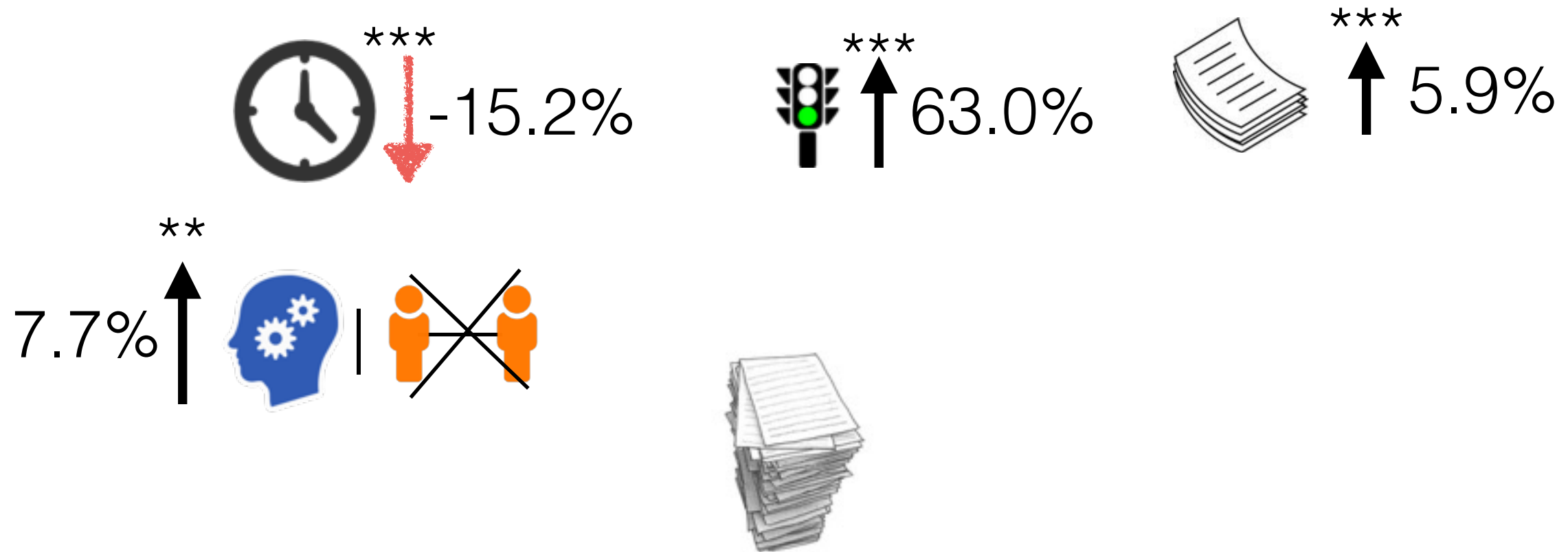


initial file  
changes





# RQ3: What are the socio-technical effects on cumulative productivity?



Negative Binomial Model

\* =  $p < 0.1$

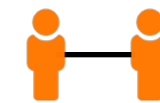
\*\* =  $p < 0.05$

\*\*\* =  $p < 0.01$

Experience



Has Links



Time period  
joined  
initial file  
changes



Is Founder



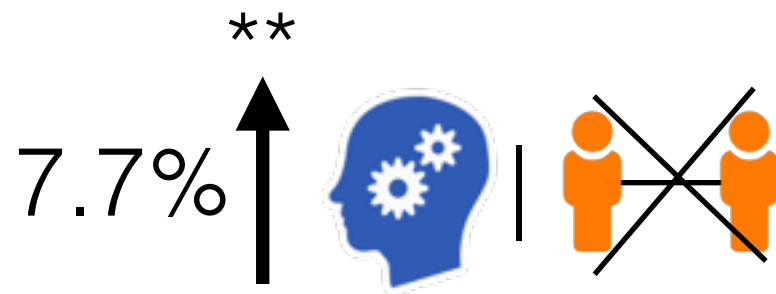
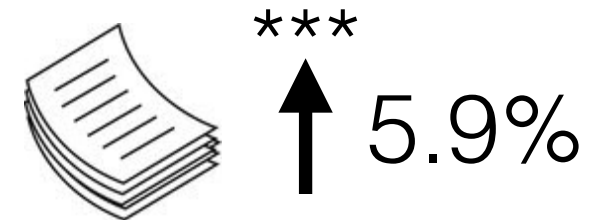
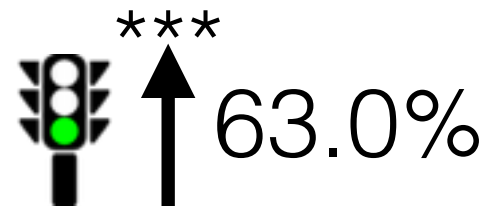
Link Strength







# RQ3: What are the socio-technical effects on cumulative productivity?



Negative Binomial Model

\* =  $p < 0.1$

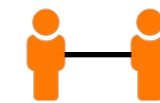
\*\* =  $p < 0.05$

\*\*\* =  $p < 0.01$

Experience



Has Links



Time period  
joined  
initial file  
changes



Is Founder

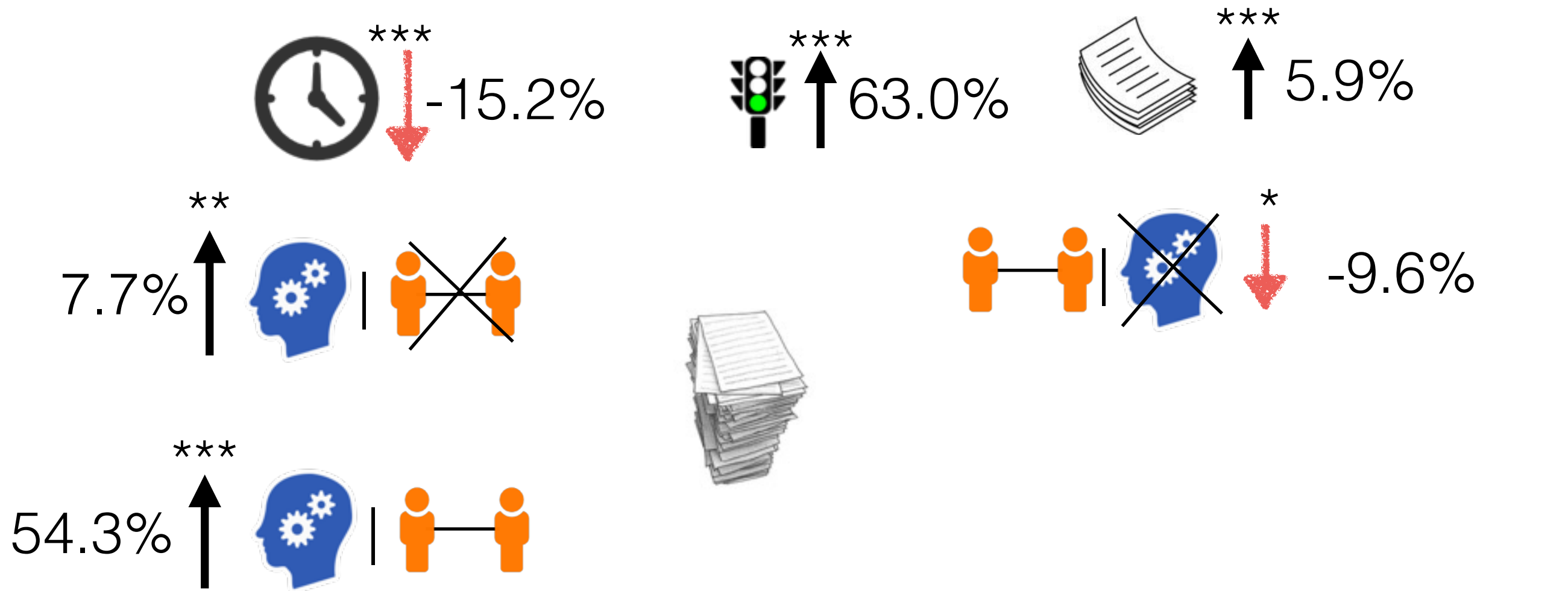


Link Strength





# RQ3: What are the socio-technical effects on cumulative productivity?



Negative Binomial Model

\* =  $p < 0.1$   
\*\* =  $p < 0.05$   
\*\*\* =  $p < 0.01$

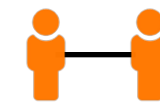
Experience



Is Founder



Has Links



Link Strength

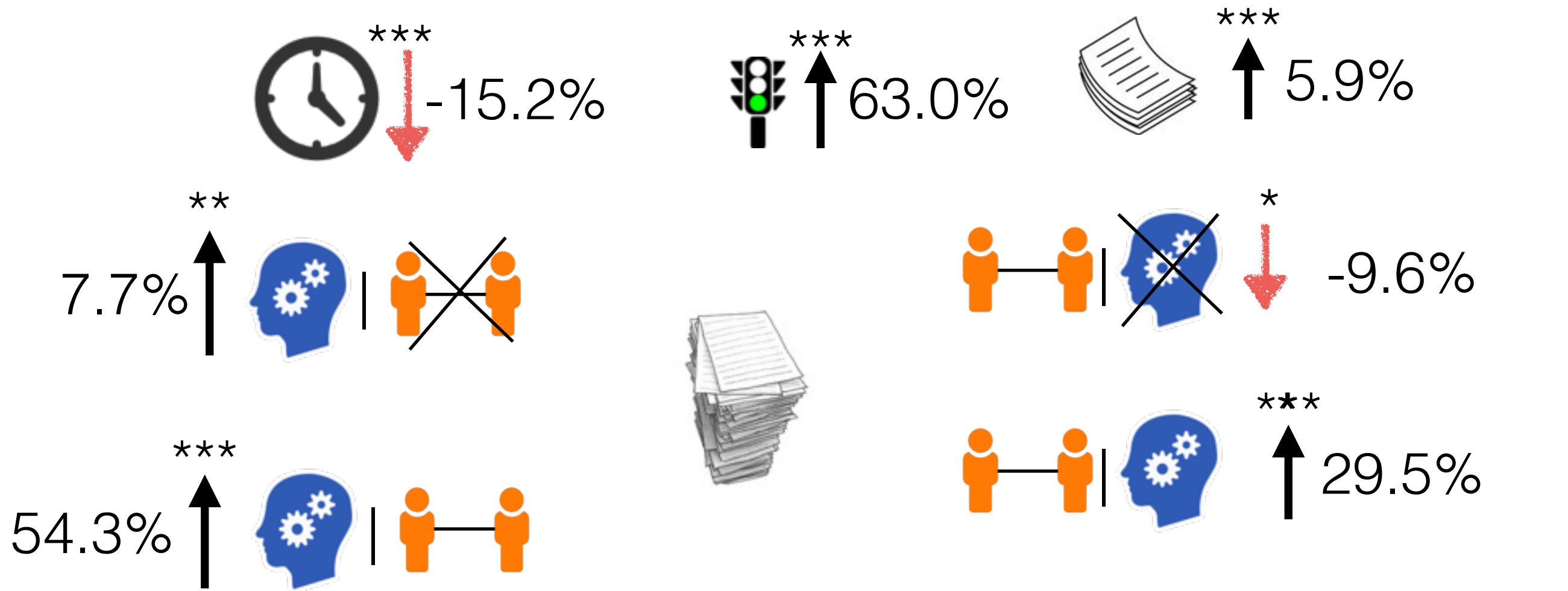


Time period  
joined  
initial file  
changes





# RQ3: What are the socio-technical effects on cumulative productivity?



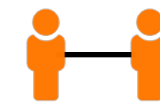
Negative Binomial Model

\* =  $p < 0.1$   
\*\* =  $p < 0.05$   
\*\*\* =  $p < 0.01$

Experience



Has Links



Time period  
joined  
initial file  
changes



Is Founder

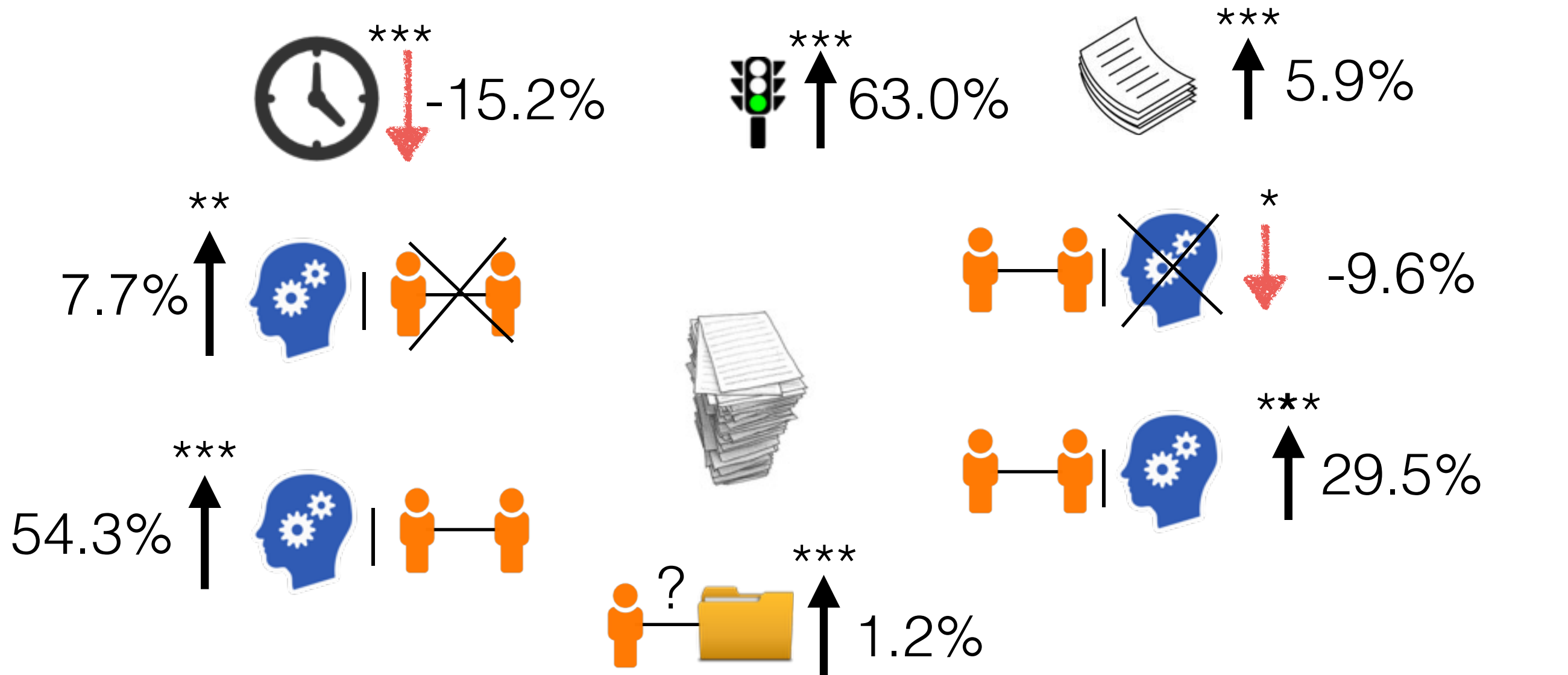


Link Strength





# RQ3: What are the socio-technical effects on cumulative productivity?



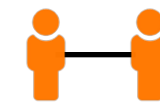
Negative Binomial Model

\* =  $p < 0.1$   
\*\* =  $p < 0.05$   
\*\*\* =  $p < 0.01$

Experience



Has Links



Time period  
joined



Is Founder



Link Strength



initial file  
changes



# Cumulative Productivity

- Having experience matters, having both social connection and experience leads to around 50% higher odds of productivity.
- The presence of a social link without experience leads to less productivity, but stronger links mitigate this.

# Conclusions + Summary

- In GitHub, developers preferentially joined projects where they have past social connections.
- Past language experience and stronger social connection better for continued contribution.
- Stronger social links helpful in the long run, but incur an initial cost.

