DALL·E 3 - “Networks of open-source software projects”

The Strength of Weak Ties in Open-Source Software Development Networks

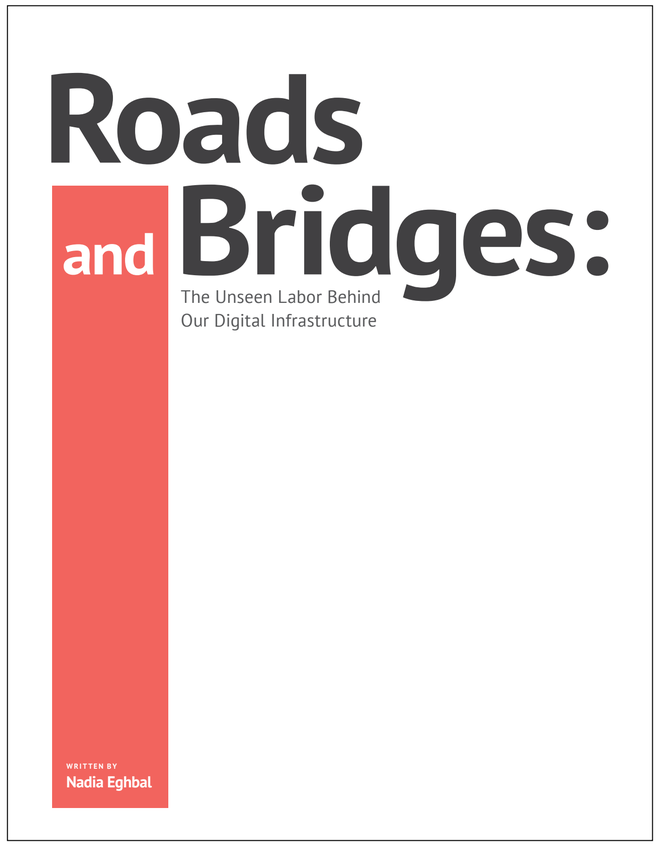
Bogdan Vasilescu
At UC Irvine, October 3rd, 2024

STRUDEL
SOCIO-TECHNICAL RESEARCH USING DATA EXCAVATION LAB

Carnegie Mellon University

S3D
Software and Societal Systems Department

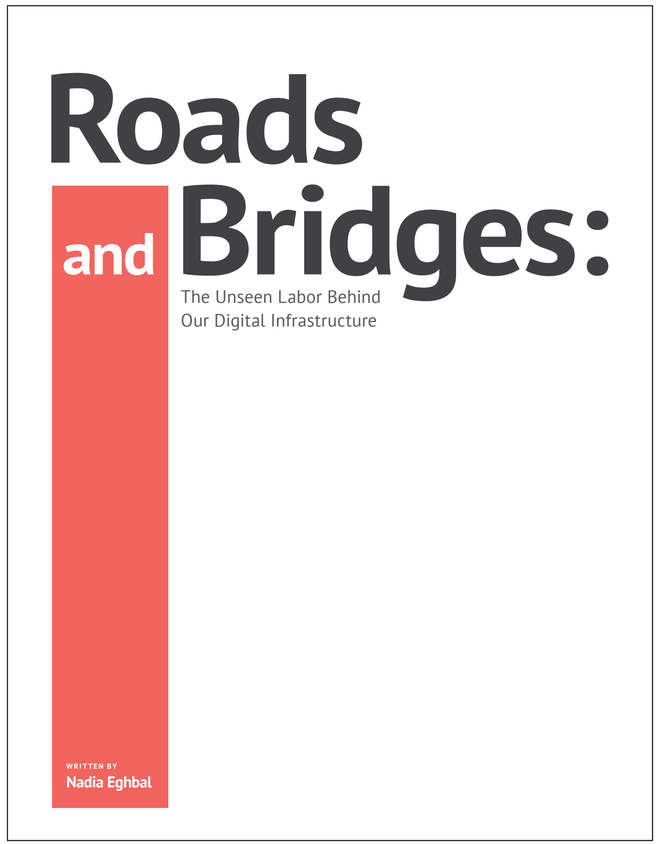
Open source software has become digital infrastructure



Everybody uses open source:

- Fortune 500 companies
- Major software companies
- Startups
- Government
- …

Like any infrastructure, it needs regular upkeep and maintenance

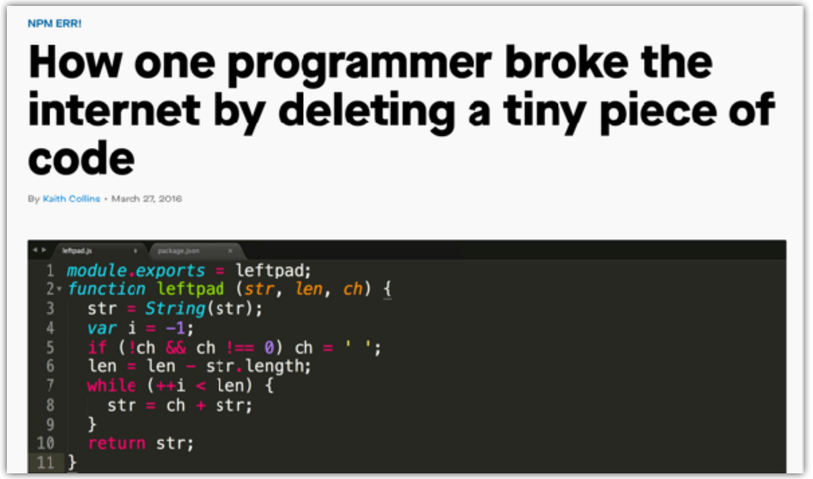


Everybody uses open source:

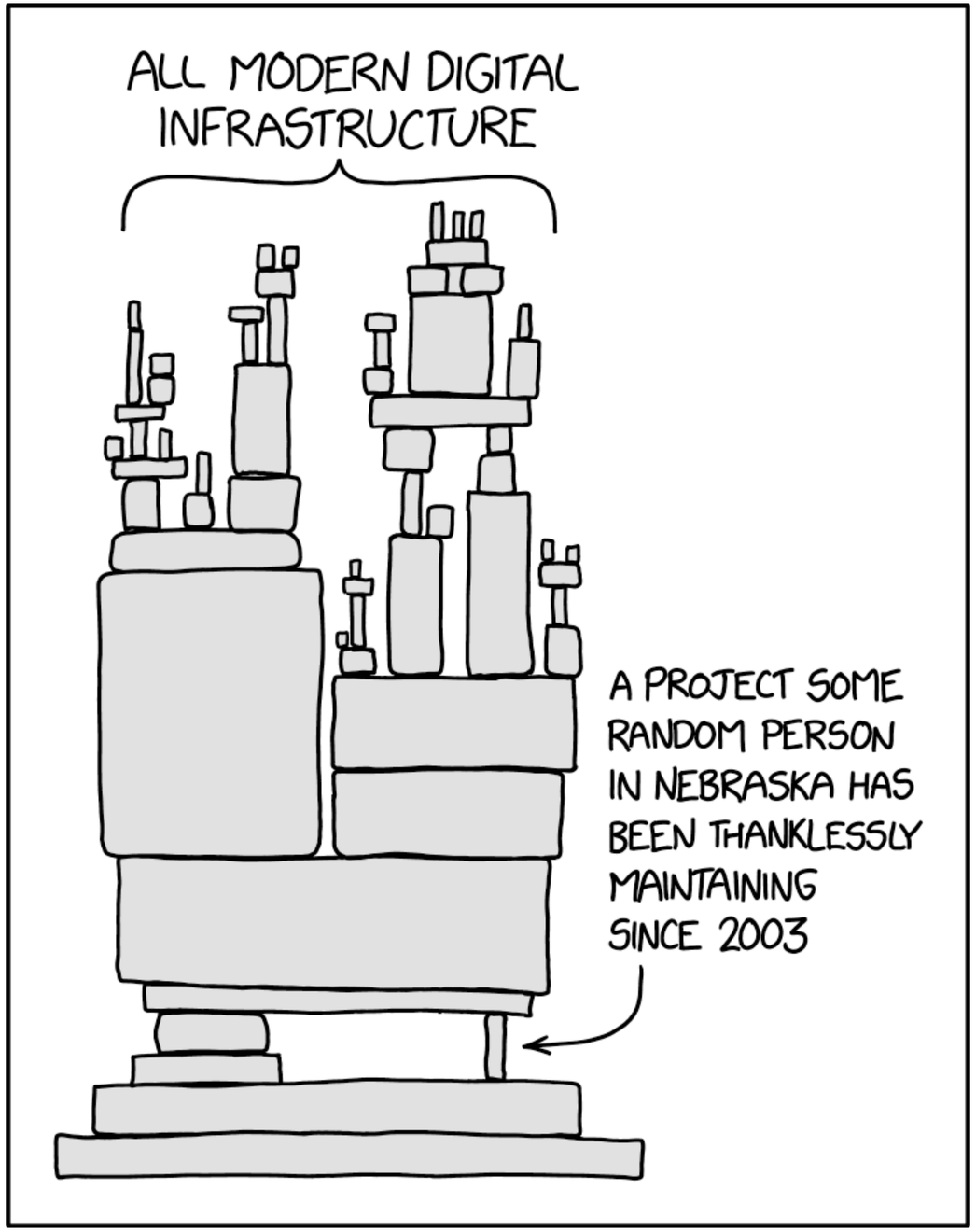
- Fortune 500 companies
- Major software companies
- Startups
- Government
- …

If undermaintained:

- Brittle supply chains
- Risks for downstream users
- Slows down innovation
- …



https://qz.com/646467/how-one-programmer-broke-the-internet-by-deleting-a-tiny-piece-of-code/



https://xkcd.com/2347/

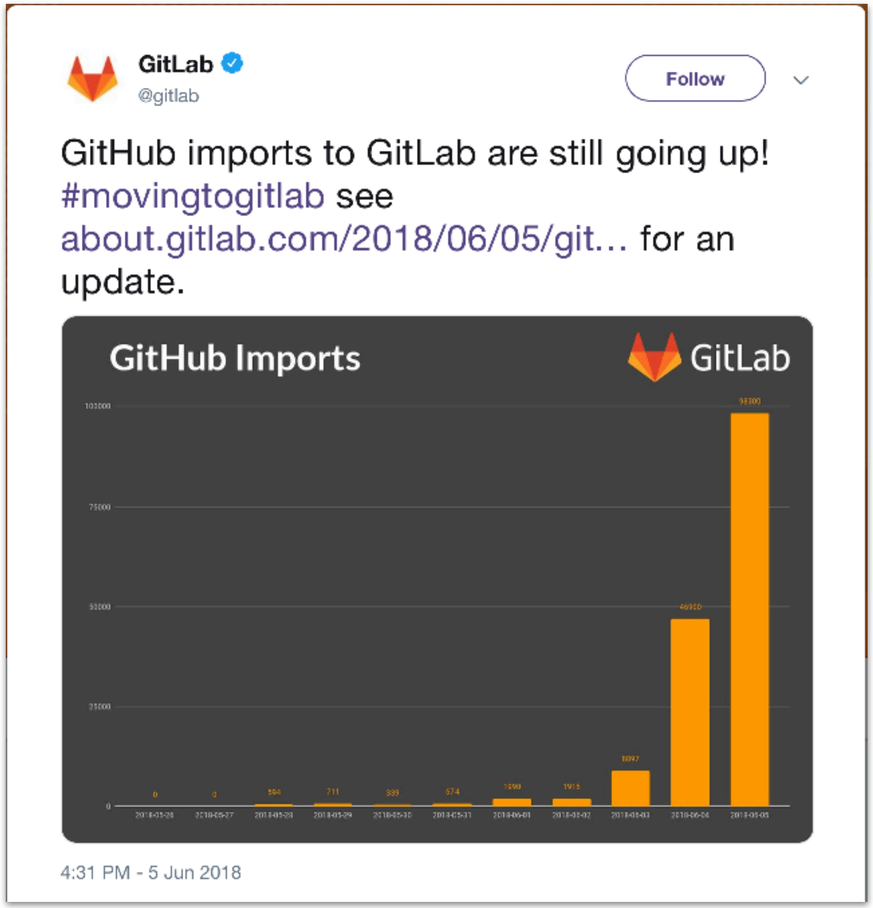
Sustaining open source is hard

Ever more open source software is being created (and reused)

Explosion of production in the past 10 years

400+ million repositories
100+ million users
(February 2024)

10+ million users
(April 2019)



There is increasing commercialization and professionalization

- Historically
  - Community-based projects (Python, RubyGems, Twisted)

- More recently, lots of commercial involvement
  - Companies (Go - Google, React - Facebook, Swift - Apple)
  - Startups (Docker, npm, Meteor)

- 23% of respondents to 2017 GitHub survey: job duties include contributing to open source

  http://opensourcesurvey.org/2017/

Expectations toward the quality, reliability, and security of open source infrastructure are high

Equifax (market cap $14 billion) built products on top of open-source infrastructure, including Apache Struts

Equifax did not make any contributions to open source projects

A flaw in Apache Struts contributed to the breach (CVE-2017-5638)

Equifax publicly blamed (with national news coverage) Apache Struts for the breach

https://www.zdnet.com/article/equifax-confirms-apache-struts-flaw-it-failed-to-patch-was-to-blame-for-data-breach/

High level of demands & stress

Easy to report issues / submit PRs

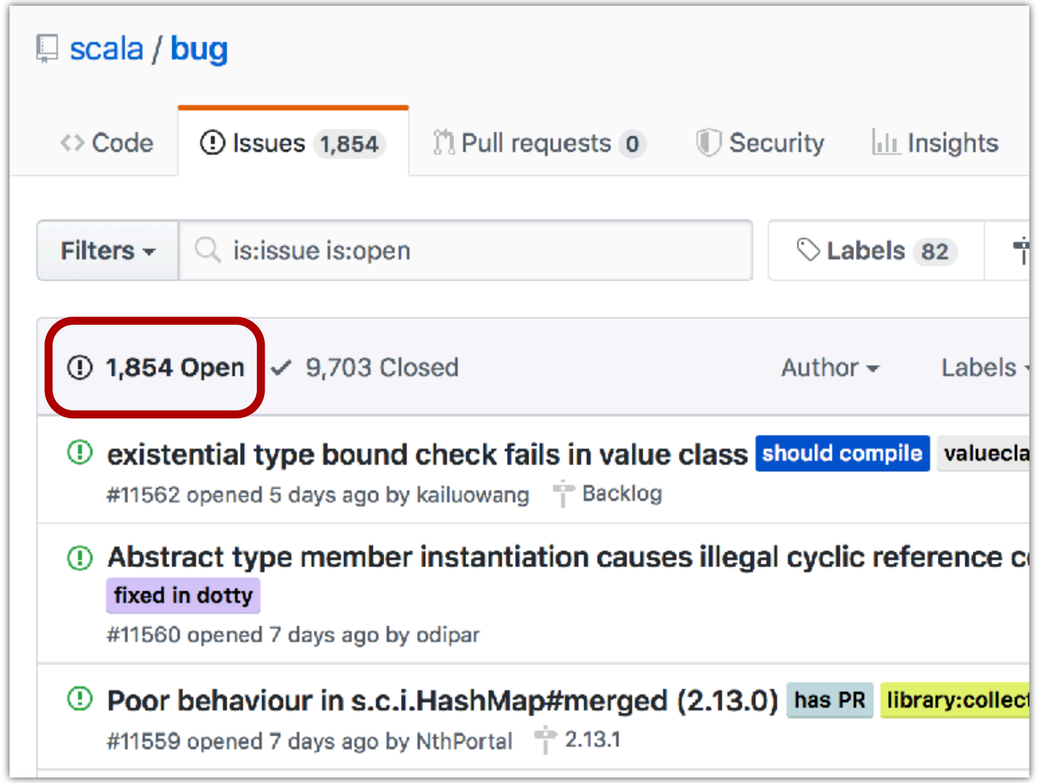
- Growing volume of requests

Social pressure to respond quickly

- Otherwise, off-putting to newcomers (Steinmacher et al. 2015)

Entitled, unreasonable users:

- *“I have been waiting 2 years for Angular to track the ‘progress’ event and it still can’t get it right?!?!”*
- *“Thank you for your ever useless explanations.”*



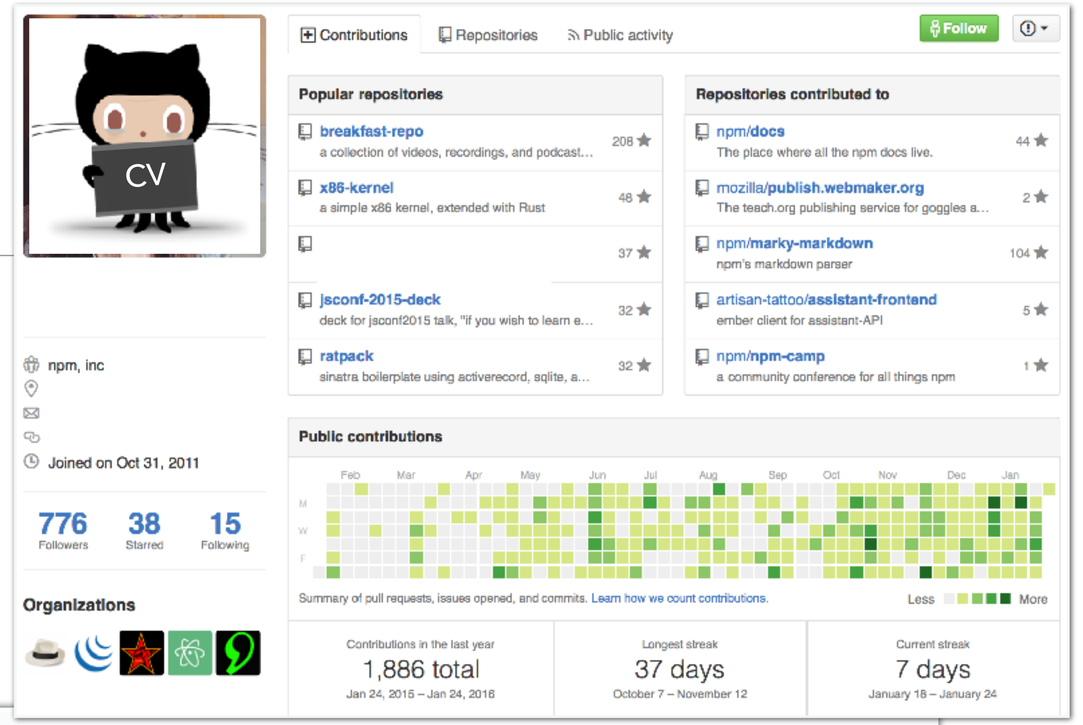
The social platforms have won

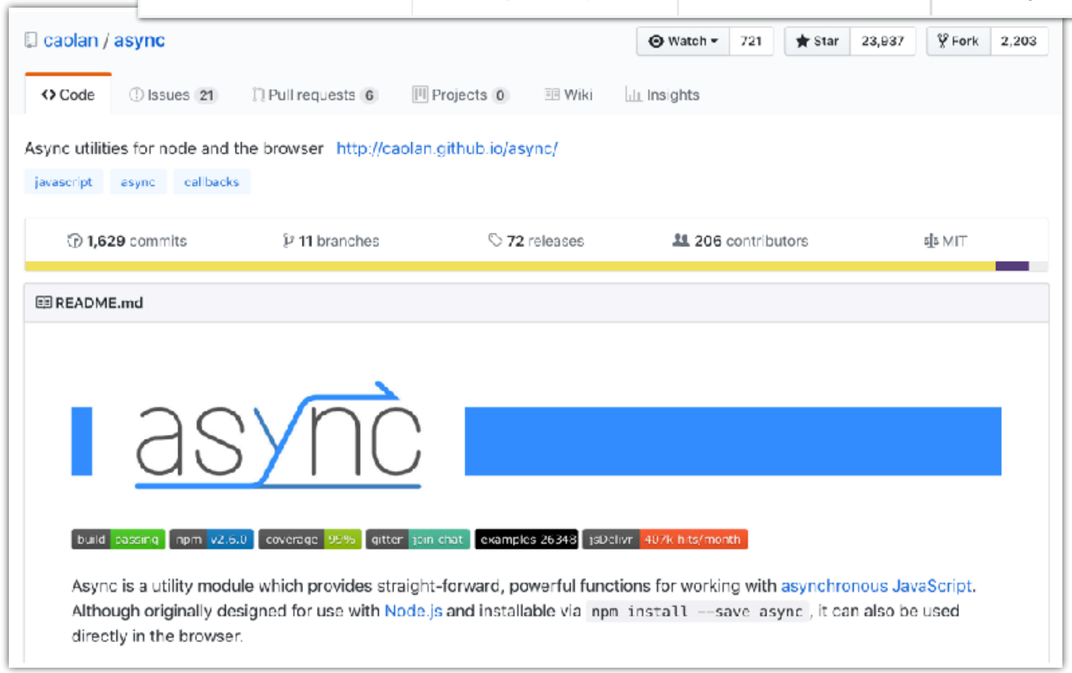
Profile pages for users and projects

Rich inferences about people's expertise and level of commitment

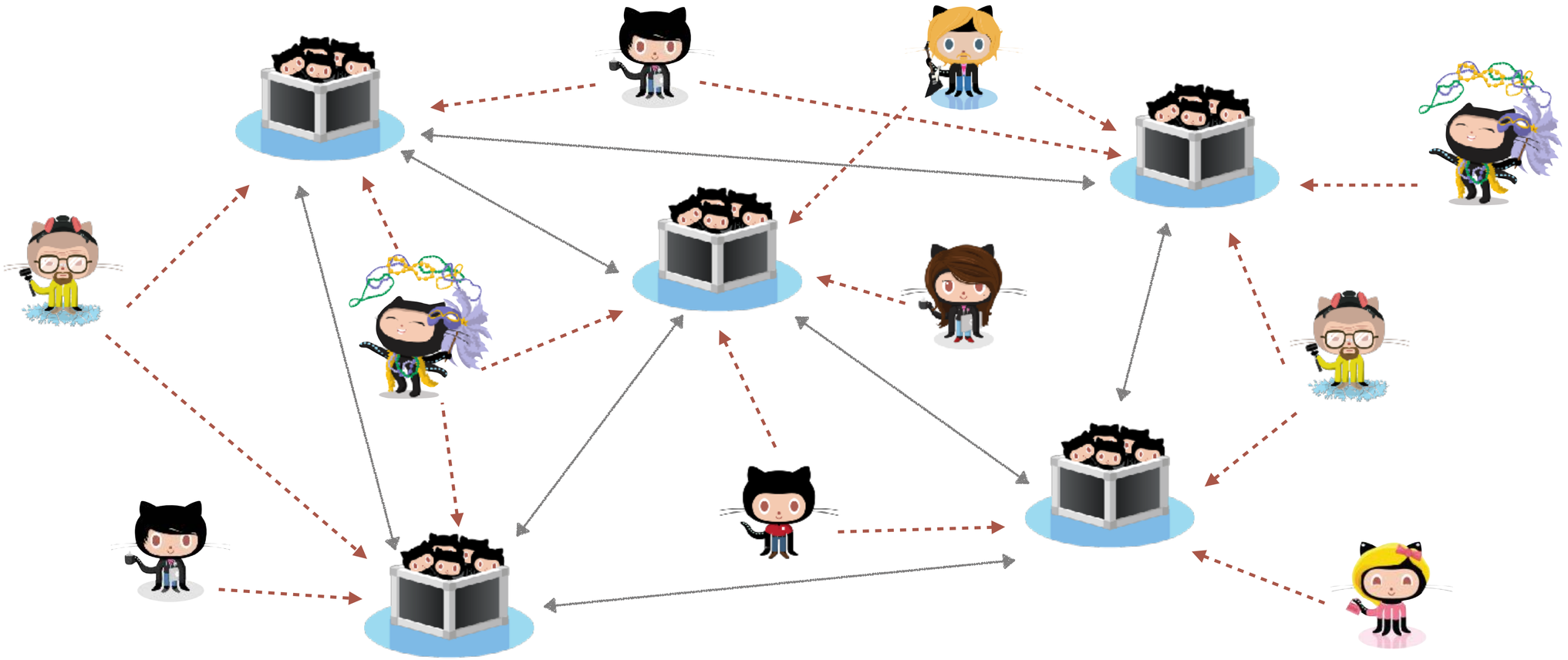
Impacts collaboration, but also recruiting and hiring

- (Dabbish et al. 2012), (Marlow et al. 2013), (Marlow and Dabbish 2013)





Contributors and projects form complex socio-technical networks!



Science is needed for evidence-based recommendations

Rich socio-technical ecosystem

Lots of changes

Lots of challenges

Little evidence or theory

Anecdotal evidence reliable? One man says "yes".

A STUDY CONDUCTED YESTERDAY by a man on himself concluded that self-reported anecdotal evidence is, in fact, both reliable and relevant.

The landmark study, conducted by Mark Mattingly of Virginia Beach in his apartment, concluded with 100% accuracy that data collected from personal experience can disprove other data conducted by reputable scientific institutions, thereby proving once and for all that "statistics can't be trusted".

In a press release Mr. Mattingly took aim at his detractors saying that "…this study shows what I've been telling people on the internet for years: all your fancy evidence and statistics don't mean nothing in the real world.".

A frequenter of internet forums, comment sections, and social media, Mr. Mattingly recounts that he was inspired to undertake the study when someone reportedly kept insisting that he provide evidence for his claims. "I think everyone's entitled to an opinion, and that my opinion is worth just as much as anyone else's" Mr. Mattingly said.

Academic types have criticised the study, and papers who are publishing it, saying that it lacks everything and makes no sense. When shown the study, Emeritus Professor James Albrecht of Carnegie Mellon University looked all confused and hopeless before making pining, guttural sounds.

Mr. Mattingly in his apartment looking all smug.

Mr. Mattingly has responded saying that this is just the first of many studies he intends to conduct, and that a meta-analysis of people who have opinions and anecdotal experiences independent of controls, methodological rigor, blinding and peer review are soon to be published, adding further weight to his initial findings.

Published Saturday 22 February 2014 by yourlogicalfallacyis.com/anecdotal *Photo: Weasello*

A great opportunity for research!

A great opportunity for research!

… because (almost) everything being archived and public makes it possible to study the problem empirically

A great opportunity for research!

… because (almost) everything being archived and public makes it possible to study the problem empirically

“The collection of public Git repositories as a whole […] exceeds 1.5PB” (Ma et al, 2021)

Ma, Y., Dey, T., Bogart, C., Amreen, S., Valiev, M., Tutko, A., ... & Mockus, A. (2021). World of code: enabling a research workflow for mining and analyzing the universe of open source VCS data. Empirical Software Engineering, 26, 1-42.

Today: Let’s look at some concrete examples of network effects

Measuring innovation
in software

Understanding how
innovation emerges

Social capital

Social contagion

STRUDEL

Today: Let’s look at some concrete examples of network effects

Measuring innovation in software

Understanding how innovation emerges

Social capital

Social contagion

Fang, Herbsleb, and Vasilescu, “Novelty Begets Long-Term Popularity, But Curbs Participation.” ICSE 2024

Open-source software development is an avenue for innovation and creative expression.

(Lakhani & Wolf, 2005)

> “How creative a person feels when working on the project is the strongest and most pervasive driver [of participation in open source]”

> “Free software is directly responsible for today’s current startup renaissance."

(Eghbal, 2016)

How to define innovation in software?

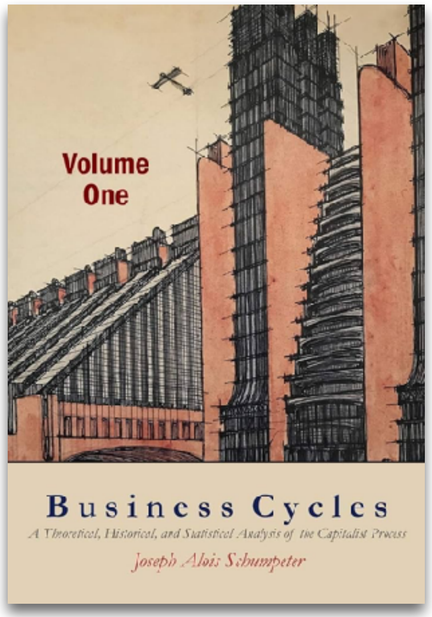
How to measure it?

How does innovation emerge?

What are its consequences?

DALL·E 3 - “An old-looking map with uncharted territory, here be dragons style”

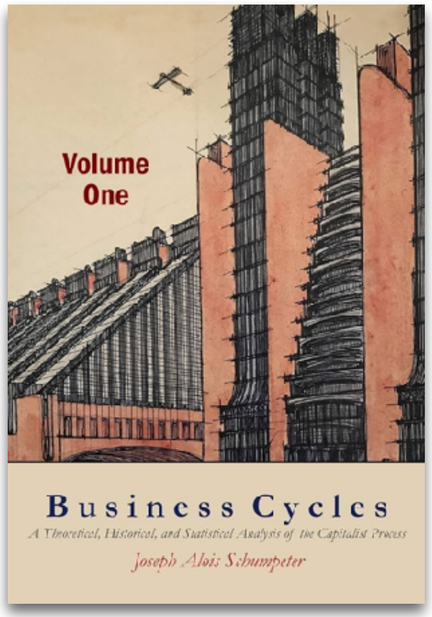
Key idea: Innovation as novel recombination



"[We may say] that innovation combines factors in a new way, or that it consists in carrying out new combinations."

(Schumpeter, 1939)

Key idea: Innovation as novel recombination



(Schumpeter, 1939)

"[We may say] that innovation combines factors in a new way, or that it consists in carrying out new combinations."

"… how scientists search for ideas is premised in part on the idea that teams can span scientific specialties, effectively combining knowledge that prompts scientific breakthroughs."

Atypical Combinations and Scientific Impact

Brian Uzzi,[1,2] Satyam Mukherjee,[1,2] Michael Stringer,[2,3] Ben Jones[1,4]*

Novelty is an essential feature of creative ideas, yet the building blocks of new ideas are often embodied in existing knowledge. From this perspective, balancing atypical knowledge with conventional knowledge may be critical to the link between innovativeness and impact. Our analysis of 17.9 million papers spanning all scientific fields suggests that science follows a nearly universal pattern: The highest-impact science is primarily grounded in exceptionally conventional combinations of prior work yet simultaneously features an intrusion of unusual combinations. Papers of this type were twice as likely to be highly cited works. Novel combinations of prior work are rare, yet teams are 37.7% more likely than solo authors to insert novel combinations into familiar knowledge domains.

Scientific enterprises are increasingly concerned that research within narrow boundaries is unlikely to be the source of the most fruitful ideas (*1*). Models of creativity emphasize that innovation is spurred through original combinations that spark new insights (*2*–*10*). Current interest in team science and how scientists search for ideas is premised in part on the idea that teams can span scientific specialties, effectively combining knowledge that prompts scientific breakthroughs (*11*–*15*).

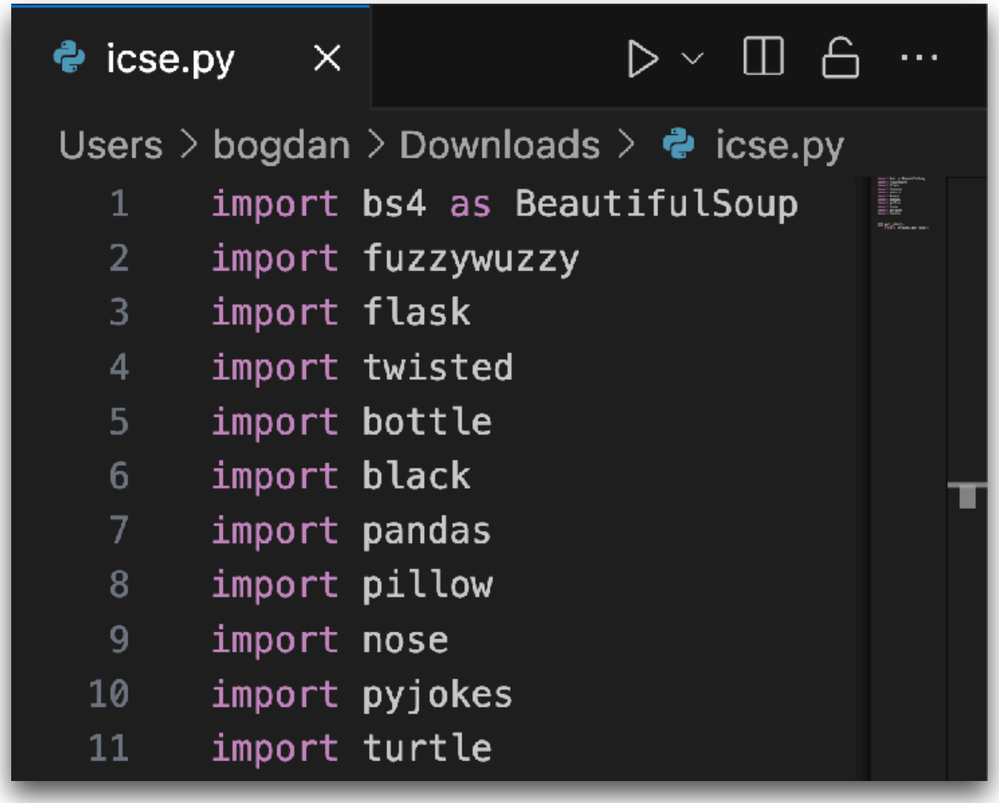
Yet the production and consumption of boundary-spanning ideas can also raise well-known challenges (*16*–*21*). If, as Einstein believed (*21*), individual scientists inevitably become narrower in their expertise as the body of scientific knowledge expands, then reaching effectively across boundaries may be increasingly challenging (*4*), especially given the difficulty of searching unfamiliar domains (*17*, *18*). Moreover, novel ideas can be difficult to absorb (*19*) and communicate, leading scientists to intentionally display conventionality. In his *Principia*, Newton presented his laws of gravitation using accepted geometry rather than his newly developed calculus, despite the latter's importance in developing his insights (*22*). Similarly, Darwin devoted the first part of the *Origin of Species* to conventional, well-accepted knowledge about the selective breeding of dogs, cattle, and birds. From this viewpoint, the balance

[1]Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, USA. [2]Northwestern Institute on Complex Systems, Northwestern University, 600 Foster, Evanston, IL 60208, USA. [3]Datascope Analytics, 180 West Adams Street, Chicago, IL 60603, USA. [4]National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138, USA.

*Corresponding author. E-mail: bjones@kellogg.northwestern.edu

(Uzzi et al, 2013)

Software innovation as novel recombination of software libraries



Lots of combinations:

- **(twisted, bottle)**
- **(turtle, nose)**
- **(black, pandas)**
- **(fuzzywuzzy, pillow)**
- …

C(n,2) unique pairs of packages.

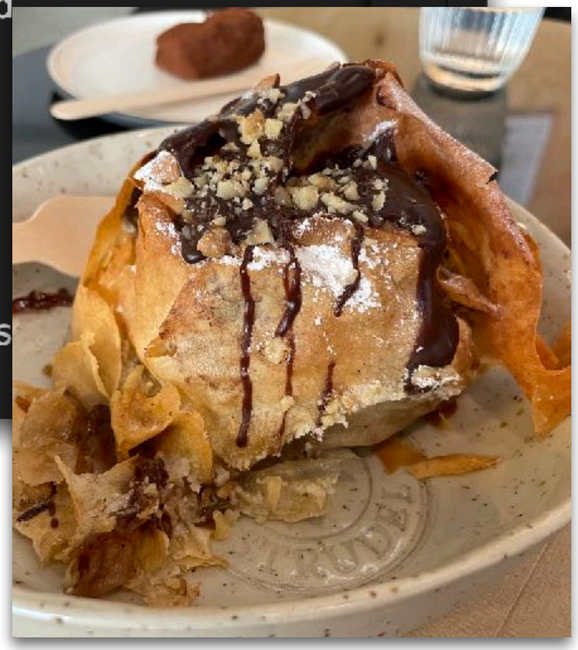
Some of these may be highly innovative because they are atypical.

Software innovation as novel recombination of software libraries

icse.py

Users > bogdan > Downloads > icse.py

```
1   import bs4 as BeautifulSoup
2   import fuzzywuzzy
3   import flask
4   import twisted
5   import bottle
6   import black
7   import pandas
8   import pillow
9   import nose
10  import pyjokes
11  import turtle
```



Lots of combinations:

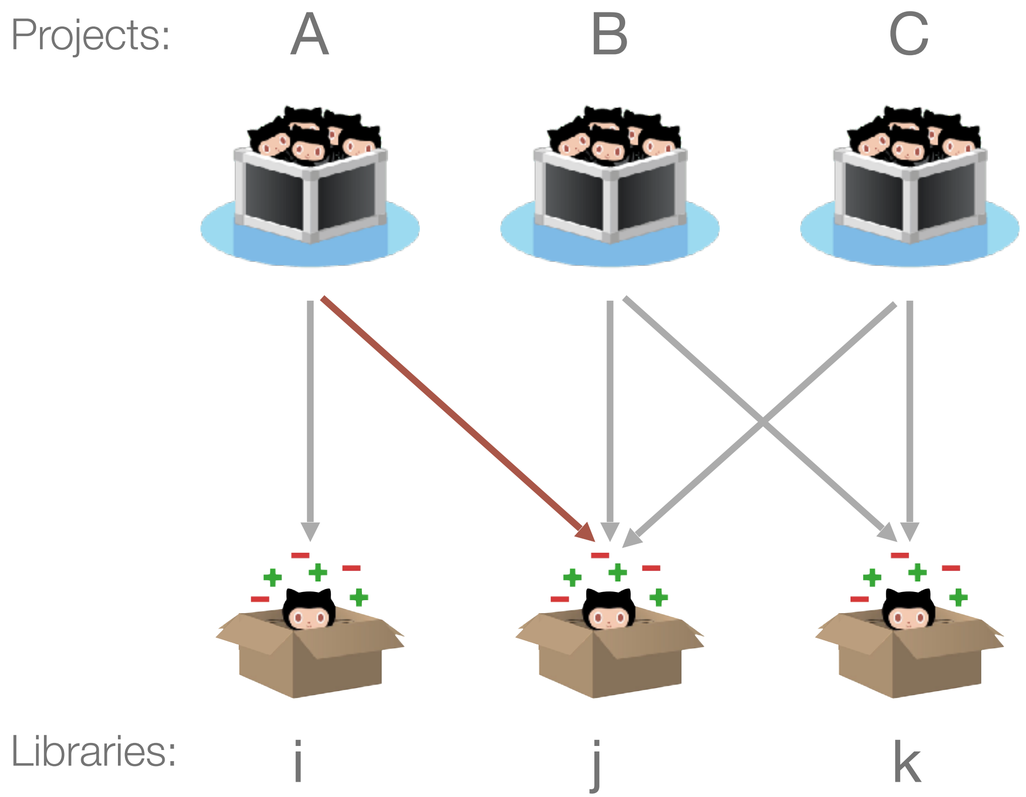
- **(twisted, bottle)**
- **(turtle, nose)**
- **(black, pandas)**
- **(fuzzywuzzy, pillow)**
- …

C(n,2) unique pairs of packages.

Dark chocolate + apple strudel is arguably innovative because it is atypical.

Key idea from network science: Comparison to null (random) model

Observed reality:

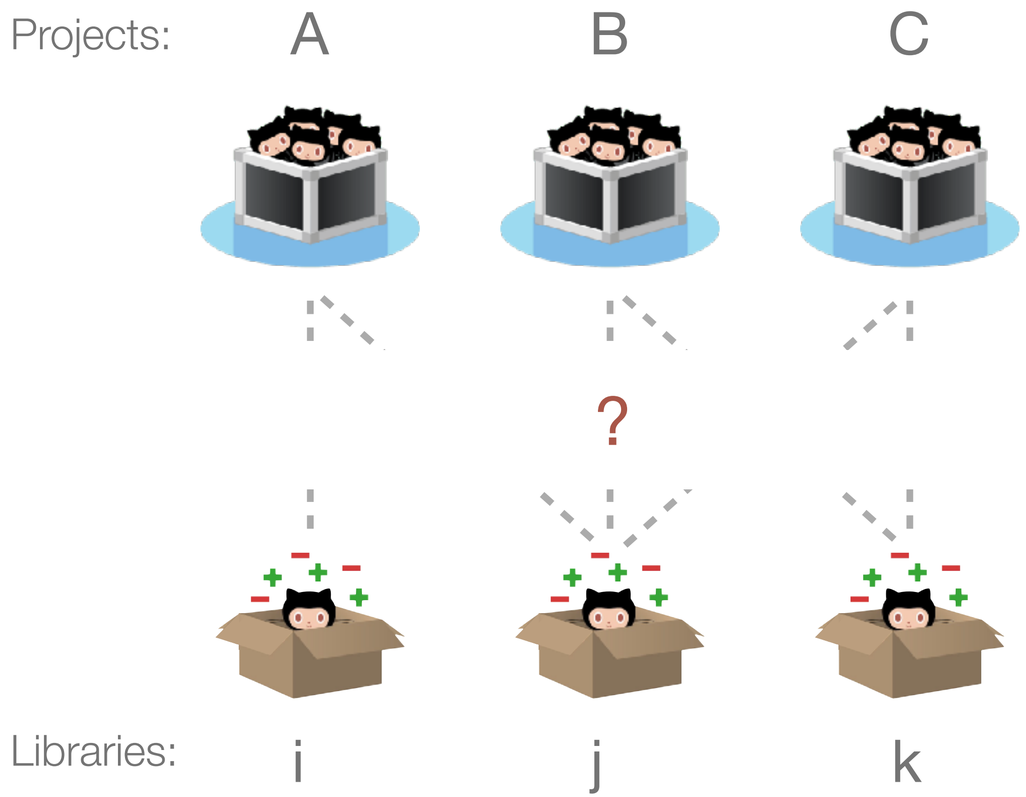


Project A adds a dependency on package j.
New combinations are formed, e.g., (i, j).

How atypical is (i, j)?

Key idea from network science: Comparison to null (random) model

Counterfactual:

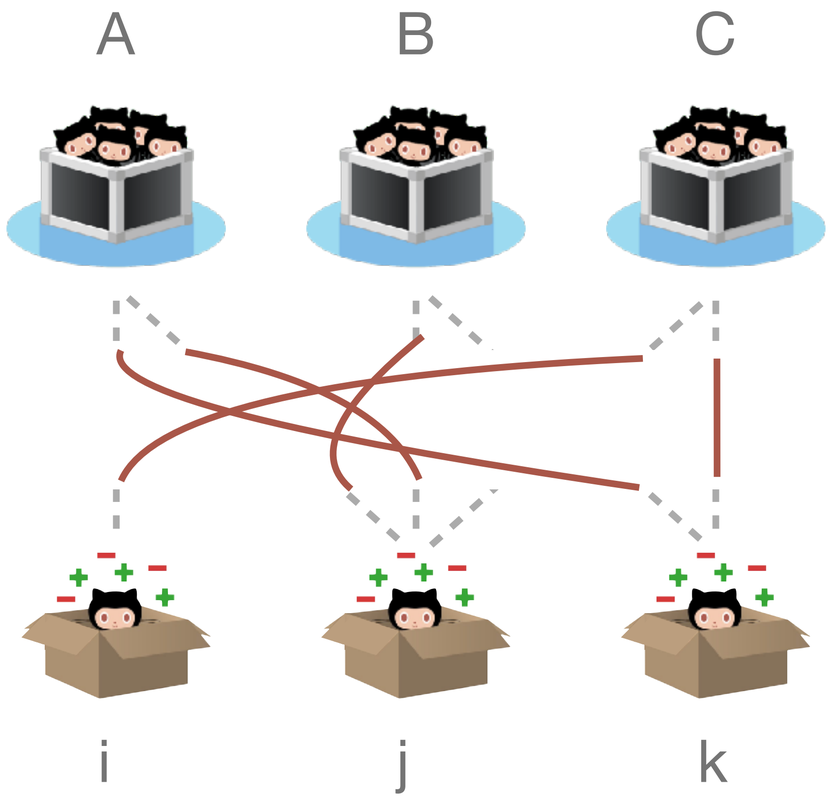


Preserve:

- all the projects
- all the libraries
- the distribution of imports per project
- the distribution of imports per library

Key idea from network science: Comparison to null (random) model

Counterfactual:



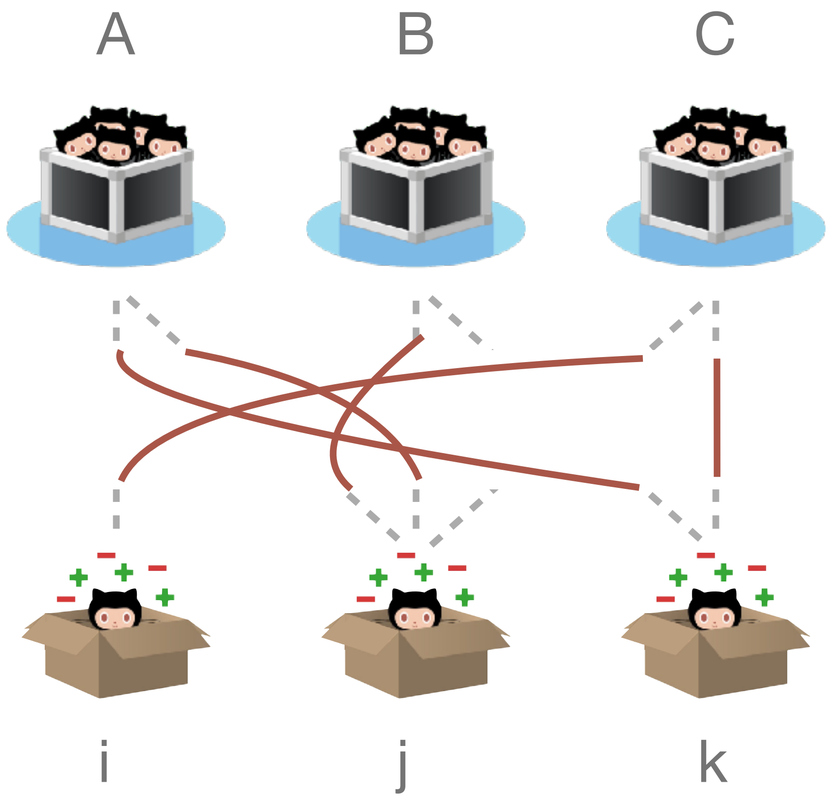
Preserve:

- all the projects
- all the libraries
- the distribution of imports per project
- the distribution of imports per library

But randomly rewire the network.

Key idea from network science: Comparison to null (random) model

Counterfactual:



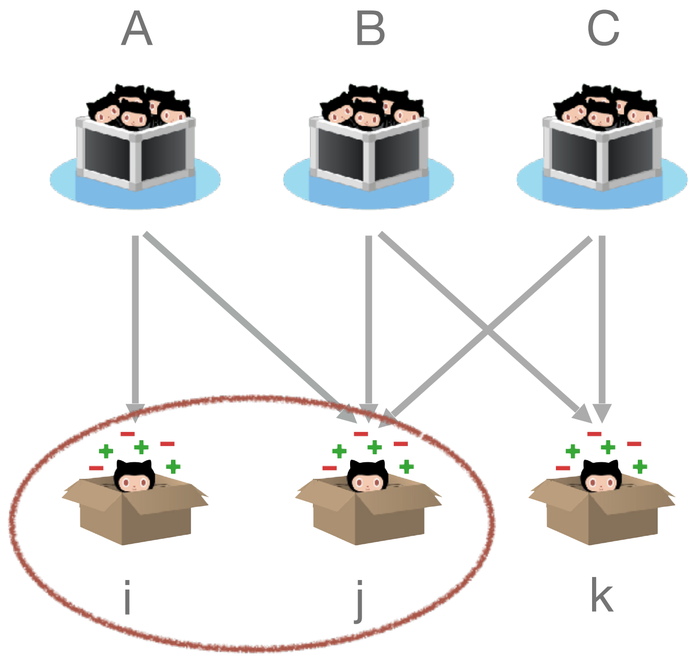
Preserve:

- all the projects
- all the libraries
- the distribution of imports per project
- the distribution of imports per library

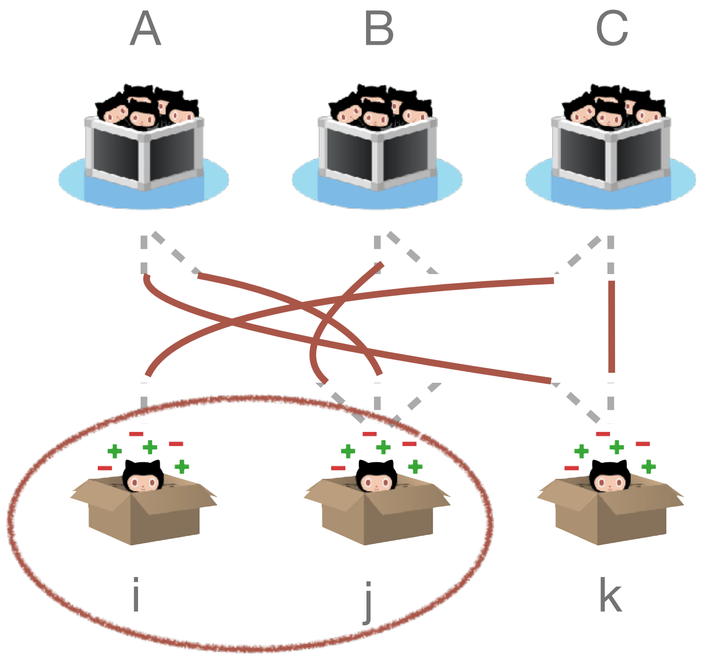
But randomly rewire the network.

And repeat many times.

This z-score estimates if two packages are used together more, less, or about as much as could be expected by chance.



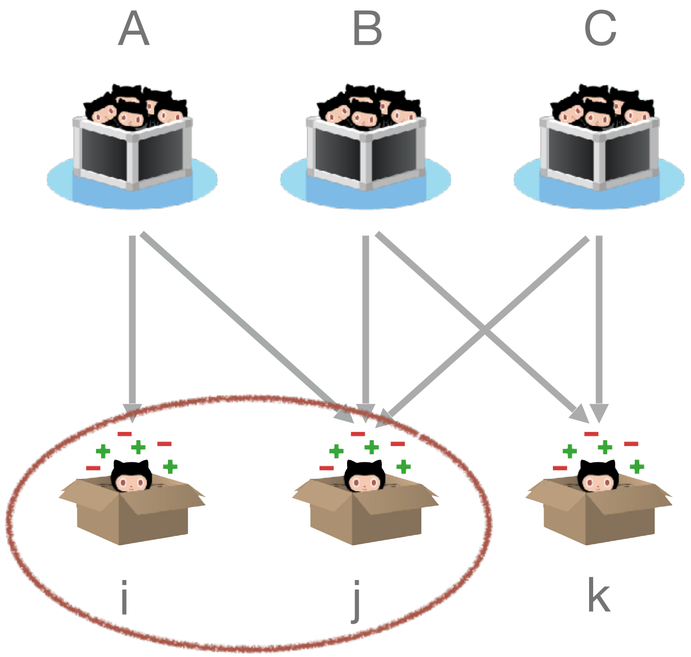
Observed number of times packages i and j appeared together until year t.



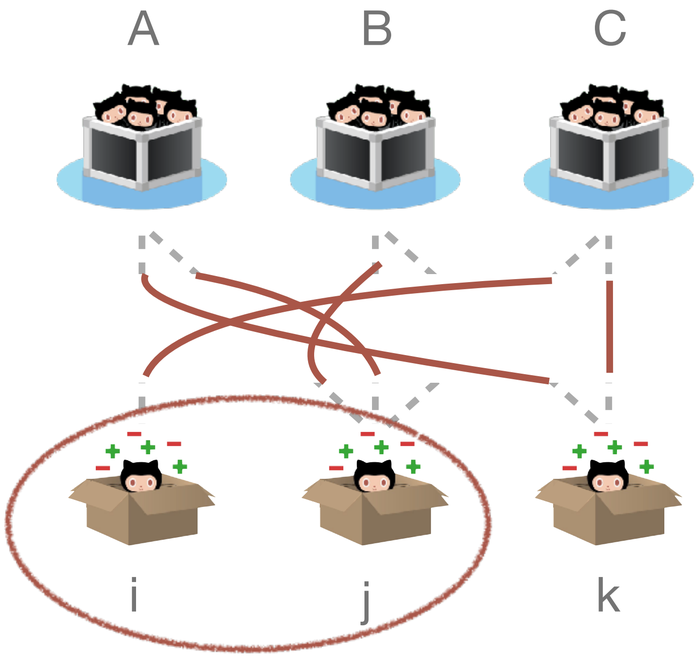
Average (i.e., expected) number of times packages i and j appeared together over N simulations.

$$z_{ijt} = (obs_{ijt} - exp_{ijt})/(\sigma_{ijt})$$

This z-score estimates if two packages are used together more, less, or about as much as could be expected by chance.



Observed number of times packages i and j appeared together until year t.



Average (i.e., expected) number of times packages i and j appeared together over N simulations.

low

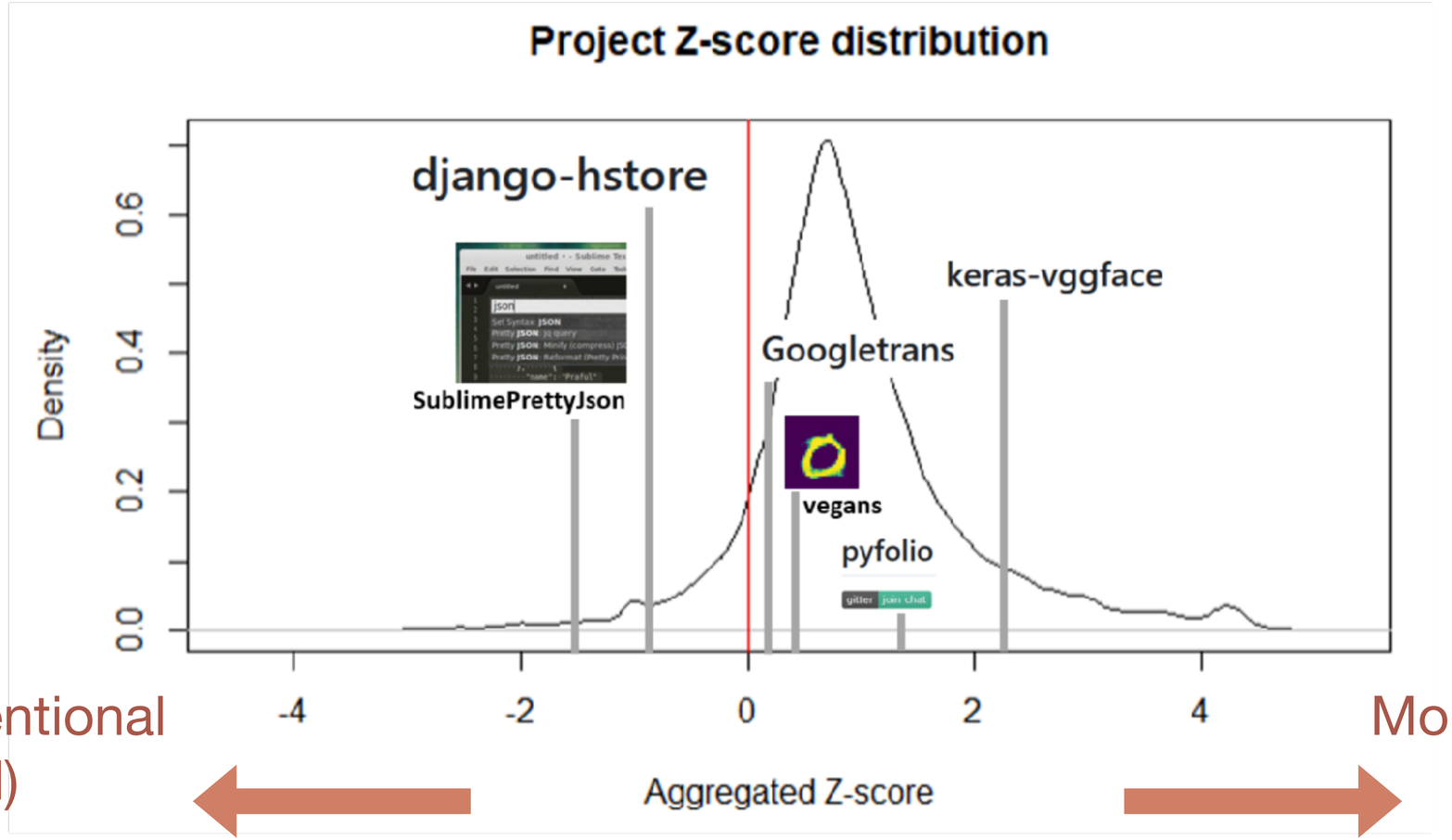
high

⇒ atypical combination

$$z_{ijt} = (obs_{ijt} - exp_{ijt})/(\sigma_{ijt})$$

Project-level aggregation is the average of pairwise atypicality z-scores

On average, projects are quite conventional.



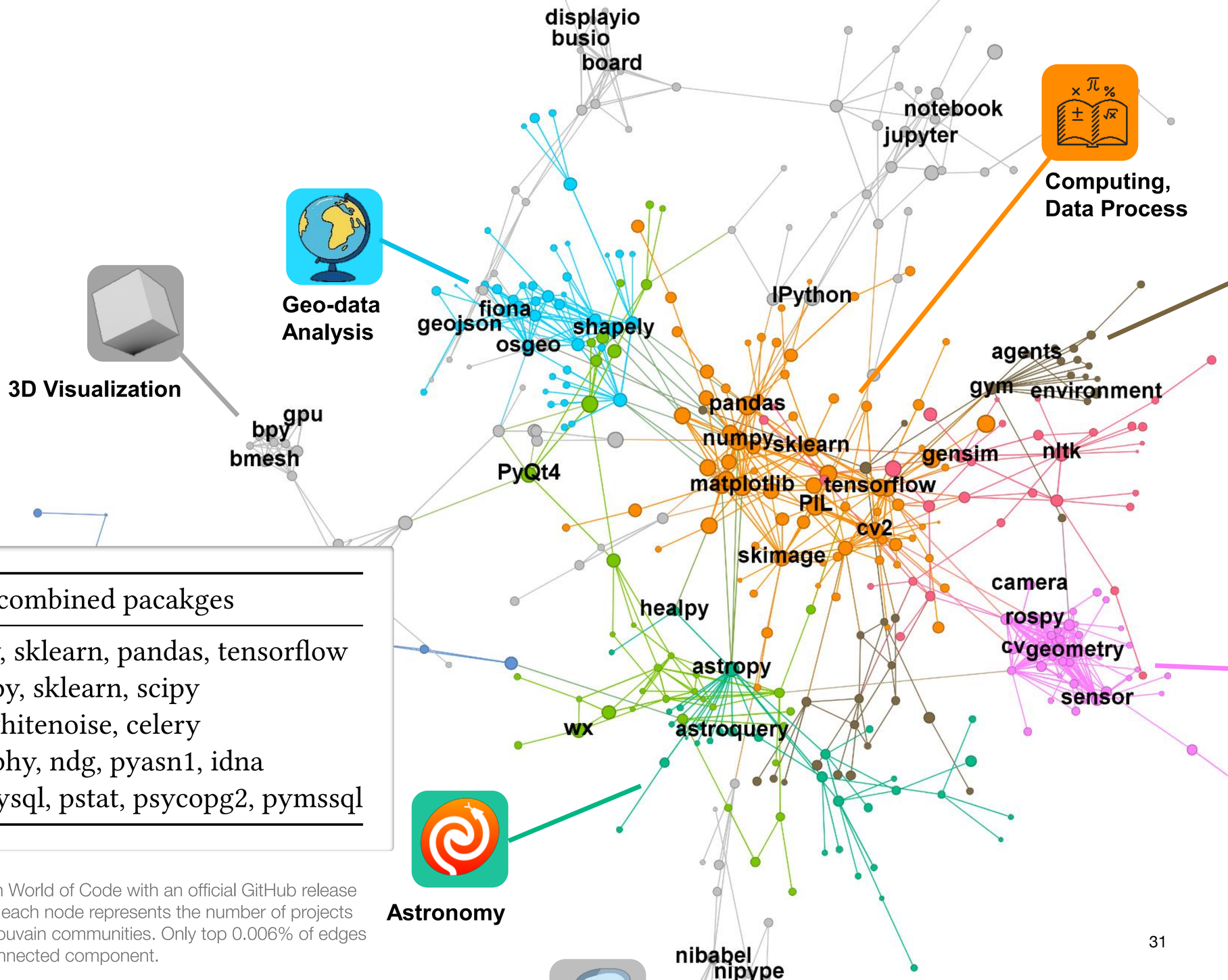
More unconventional (atypical)

More conventional (typical)

Sanity checking

No ground truth on atypical package combinations, but at least the typical combinations should be meaningful!

Focal package	Top five mostly combined pacakges
numpy	matplotlib, scipy, sklearn, pandas, tensorflow
tensorflow	keras, cv2, numpy, sklearn, scipy
django	rest, dj, south, whitenoise, celery
OpenSSL	ntlm, cryptography, ndg, pyasn1, idna
pymysql	MySQLdb, aiomysql, pstat, psycopg2, pymssql



Fine print: Starting data consists of all Python projects in World of Code with an official GitHub release (75,388 projects and 7,728 packages total). The size of each node represents the number of projects that imported the package by 2019. Colors represent Louvain communities. Only top 0.006% of edges with the highest z-score shown, and only the largest connected component.

Software innovation as novel recombination of software libraries

Combining software libraries that are not often used together is like using unusual ingredients in your cooking.

- People may be impressed by your culinary creativity.
- Serving unusual dishes can be risky if the chefs are unable to perfect the recipes and the customers are unwilling to try new things.

https://www.tasteofhome.com/recipes/chocolate-peanut-butter-pizza/

Software innovation as novel recombination of software libraries

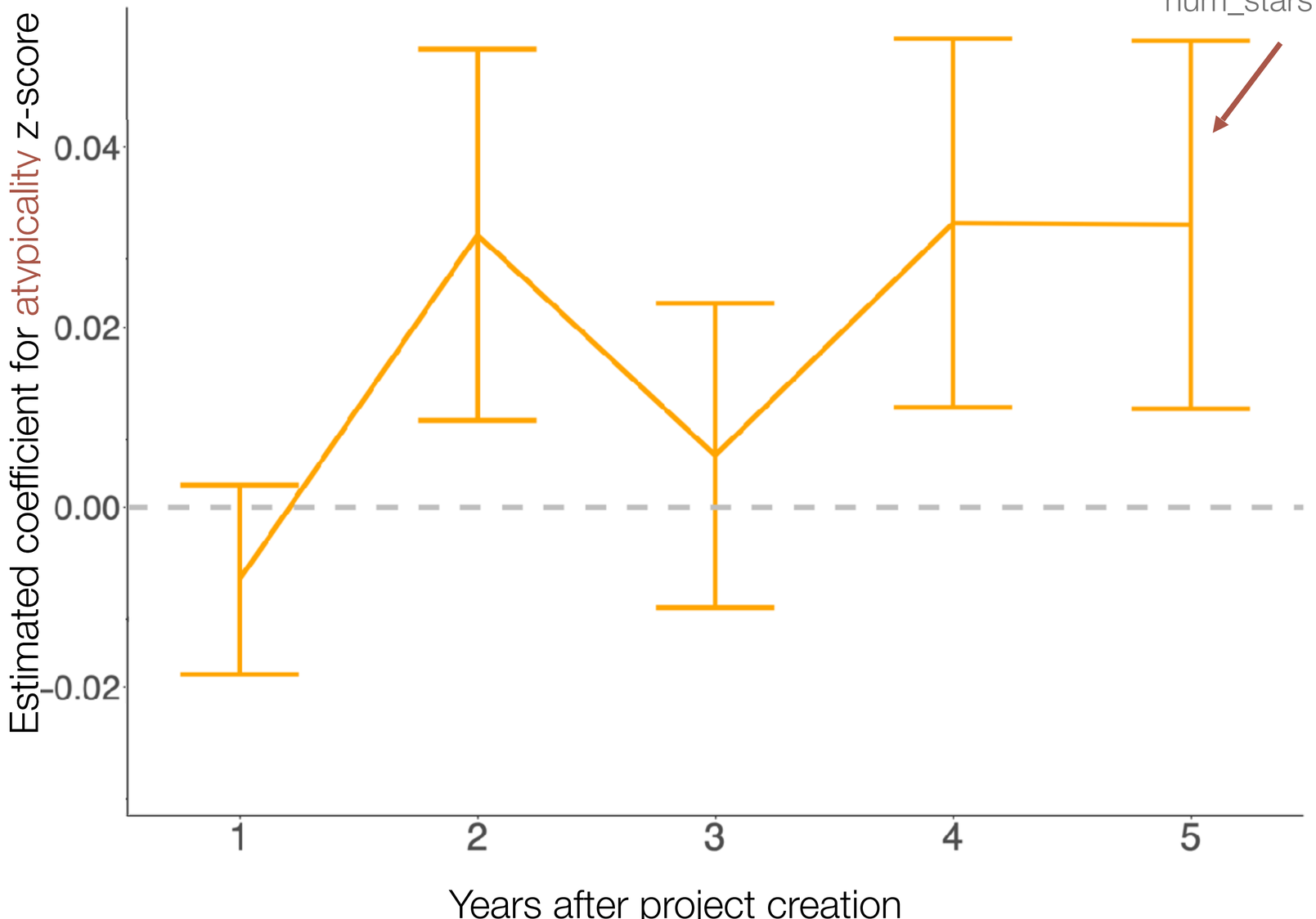
Combining software libraries that are not often used together is like using unusual ingredients in your cooking.

- Hyp: Projects that use more atypical combinations of libraries tend to be more popular.
- Hyp: More innovative projects tend to be less sustainable in the long term.

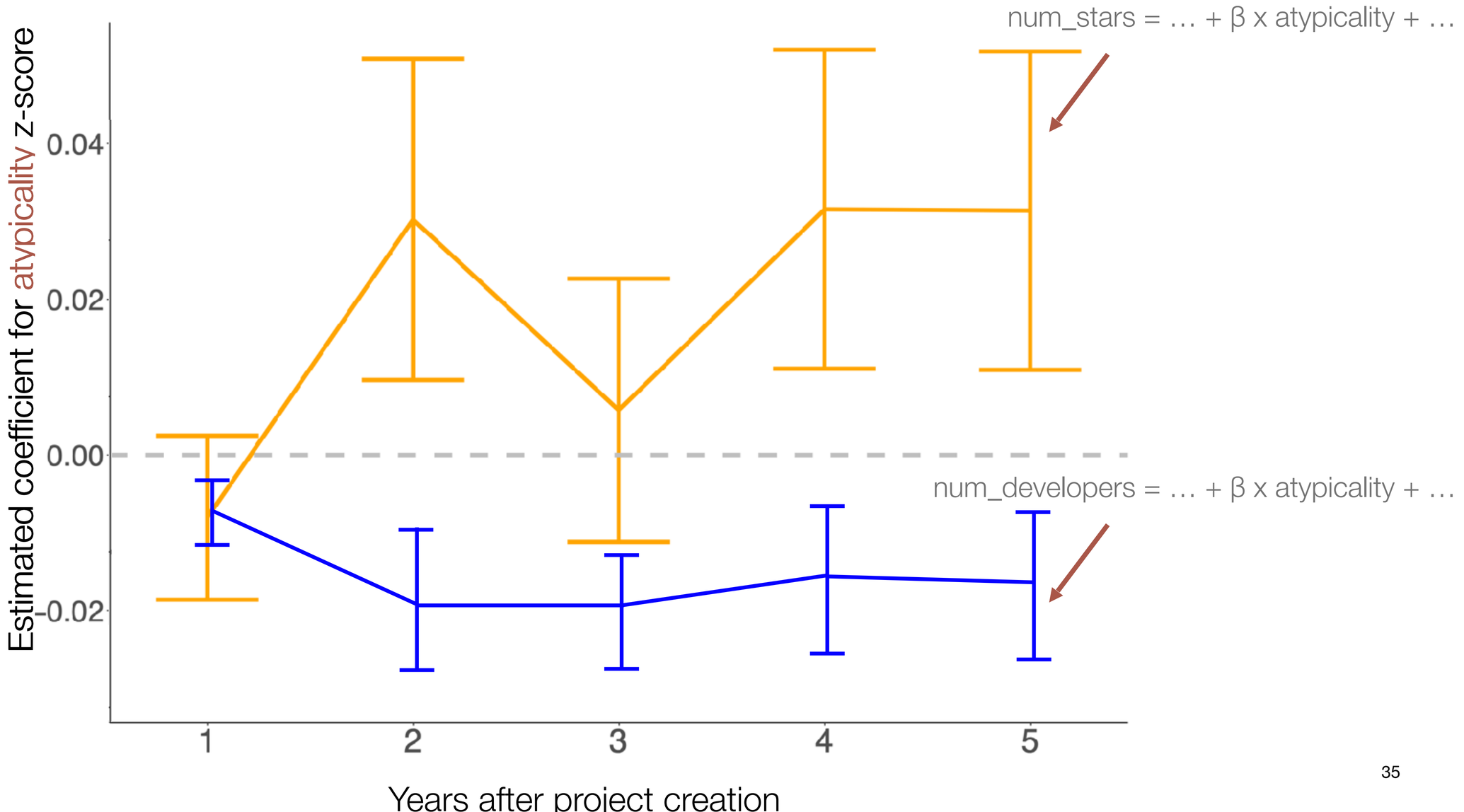
https://www.tasteofhome.com/recipes/chocolate-peanut-butter-pizza/

Atypical (novel) projects tend to have more stars.

num_stars = … + β x atypicality + …

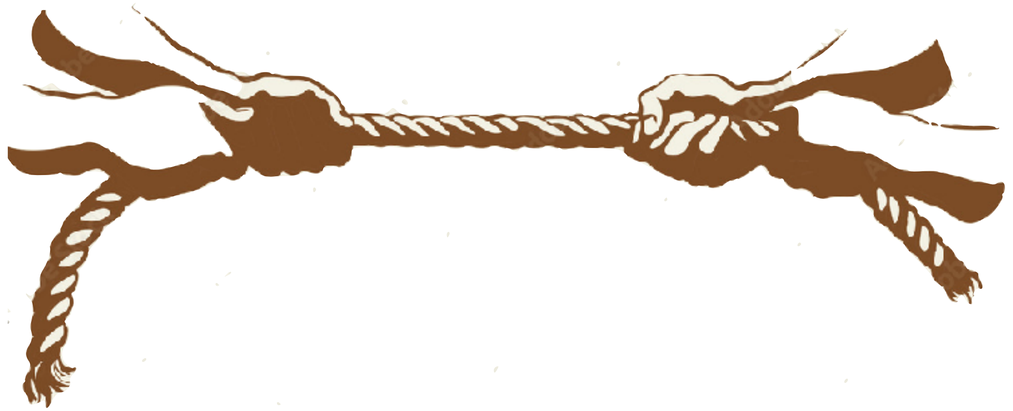


Atypical (novel) projects tend to have smaller teams (and higher probability of becoming abandoned).



Tension between innovation and open source sustainability?

Incentive to create ever-new things



The “grunt work” of maintaining existing systems

- Creative expression is a main driver of contributing to open source
- Innovation seems to be rewarded with increased popularity

Will it become increasingly harder to ensure that sufficient maintenance attention (developers, funding, etc) is being allocated to the projects that need it the most?

Today: Let’s look at some concrete examples of network effects

Measuring innovation in software

Understanding how innovation emerges

Social capital

Social contagion

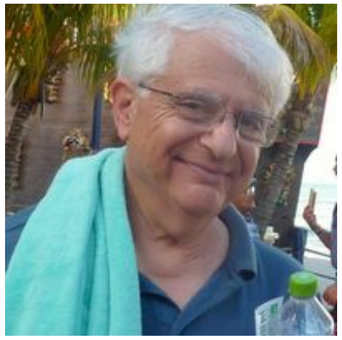
Fang, Park, Evans, Herbsleb, and Vasilescu, “Weak Ties Explain Open Source Innovation.” Under review

Once upon a time, a PhD student at Harvard University was writing their dissertation …

Stanford
Sociology
SCHOOL OF HUMANITIES AND SCIENCES

Mark Granovetter

Joan Butler Ford Professor in the School of Humanities and Sciences; Professor of Sociology



A.B. Princeton University 1965
Modern European and American History
Ph.D. Harvard University 1970
Sociology

https://sociology.stanford.edu/people/mark-granovetter

The Strength of Weak Ties[1]

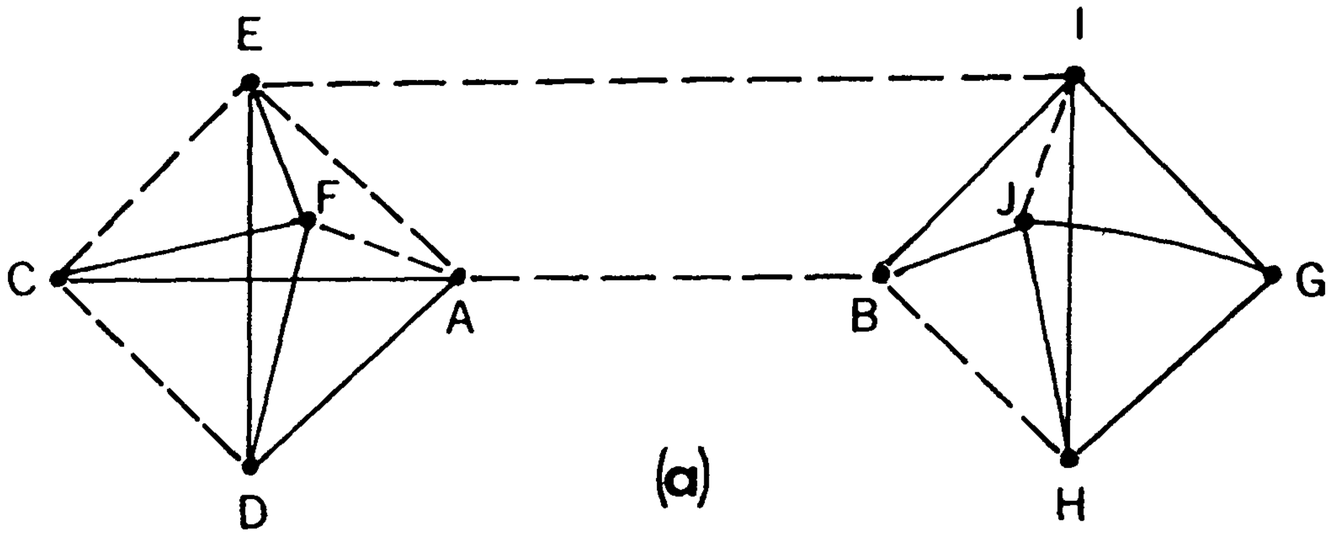
Mark S. Granovetter
Johns Hopkins University

Analysis of social networks is suggested as a tool for linking micro and macro levels of sociological theory. The procedure is illustrated by elaboration of the macro implications of one aspect of small-scale interaction: the strength of dyadic ties. It is argued that the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another. The impact of this principle on diffusion of influence and information, mobility opportunity, and community organization is explored. Stress is laid on the cohesive power of weak ties. Most network models deal, implicitly, with strong ties, thus confining their applicability to small, well-defined groups. Emphasis on weak ties lends itself to discussion of relations *between* groups and to analysis of segments of social structure not easily defined in terms of primary groups.

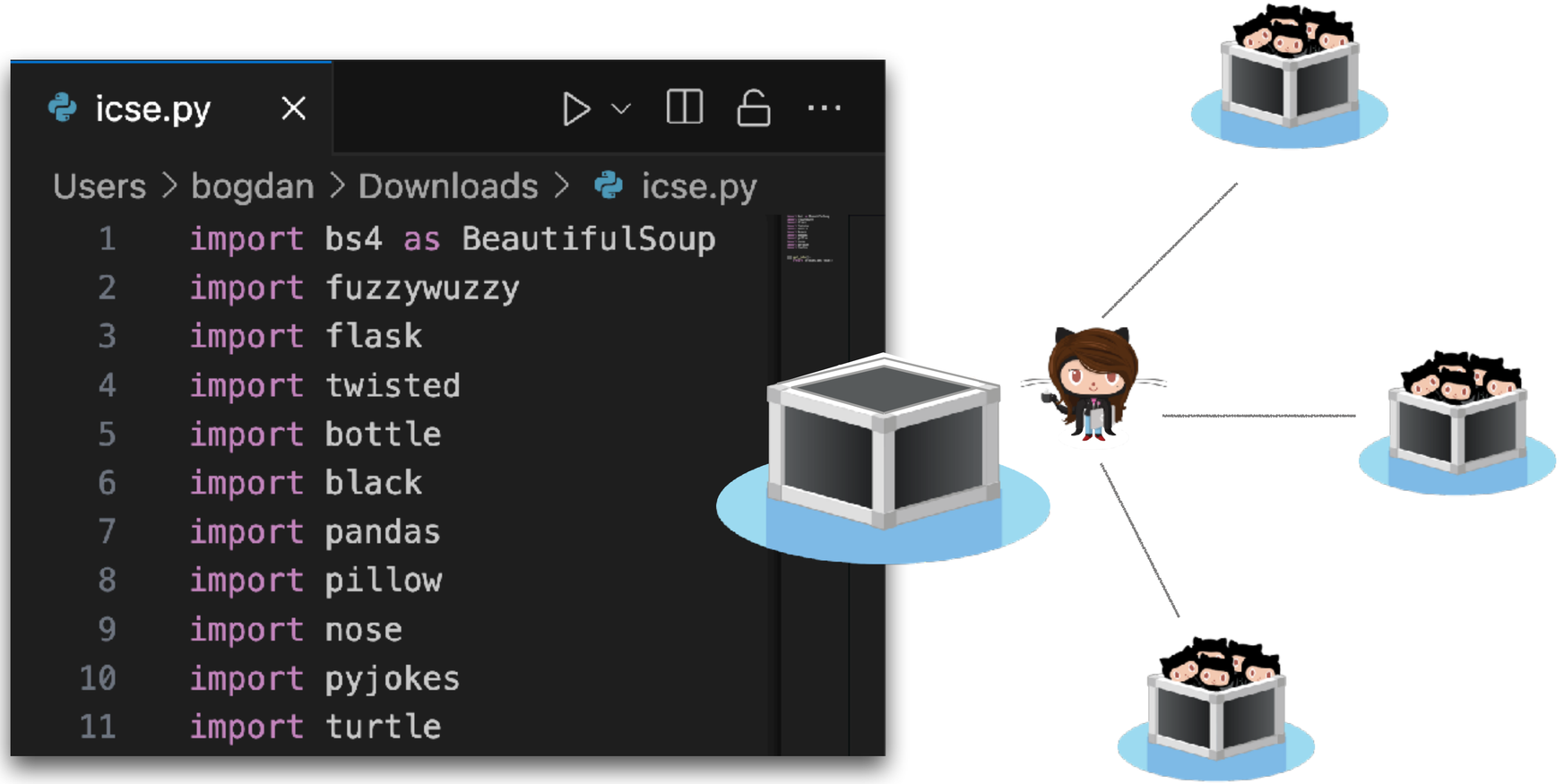
Weak ties are more effective in job searches because they act as bridges.

The majority of people found their jobs through acquaintances (weak ties) rather than close friends or family (strong ties).

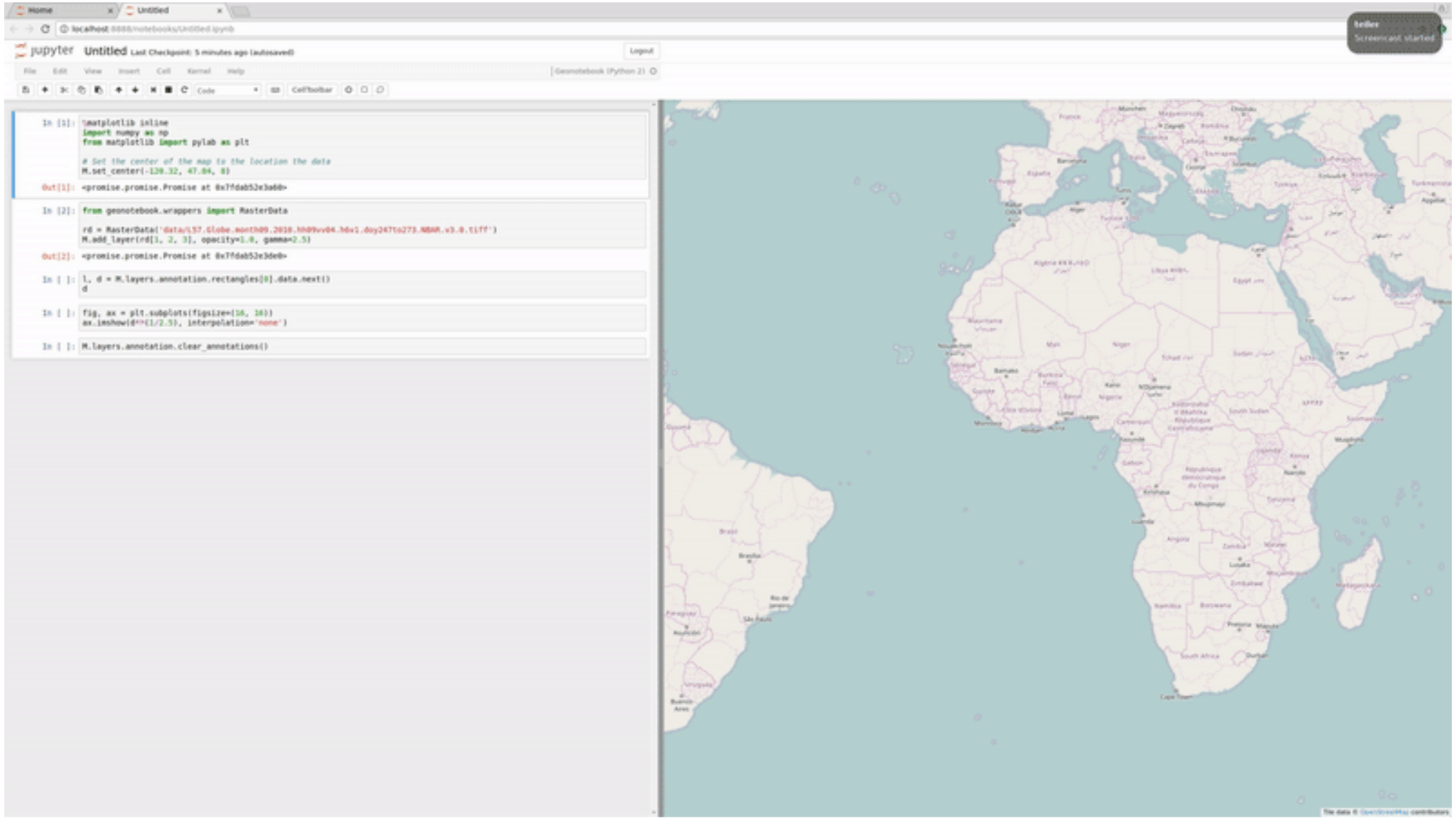
In a random sample of recent professional, technical, and managerial job changers living in a Boston suburb, I asked those who found a new job through contacts how often they *saw* the contact around the time that he passed on job information to them. I will use this as a measure of tie strength.[15] A natural a priori idea is that those with whom one has strong ties are more motivated to help with job information. Opposed to this greater motivation are the structural arguments I have been making: those to whom we are weakly tied are more likely to move in circles different from our own and will thus have access to information different from that which we receive.



Do OSS developers also find their new ideas through weak ties?

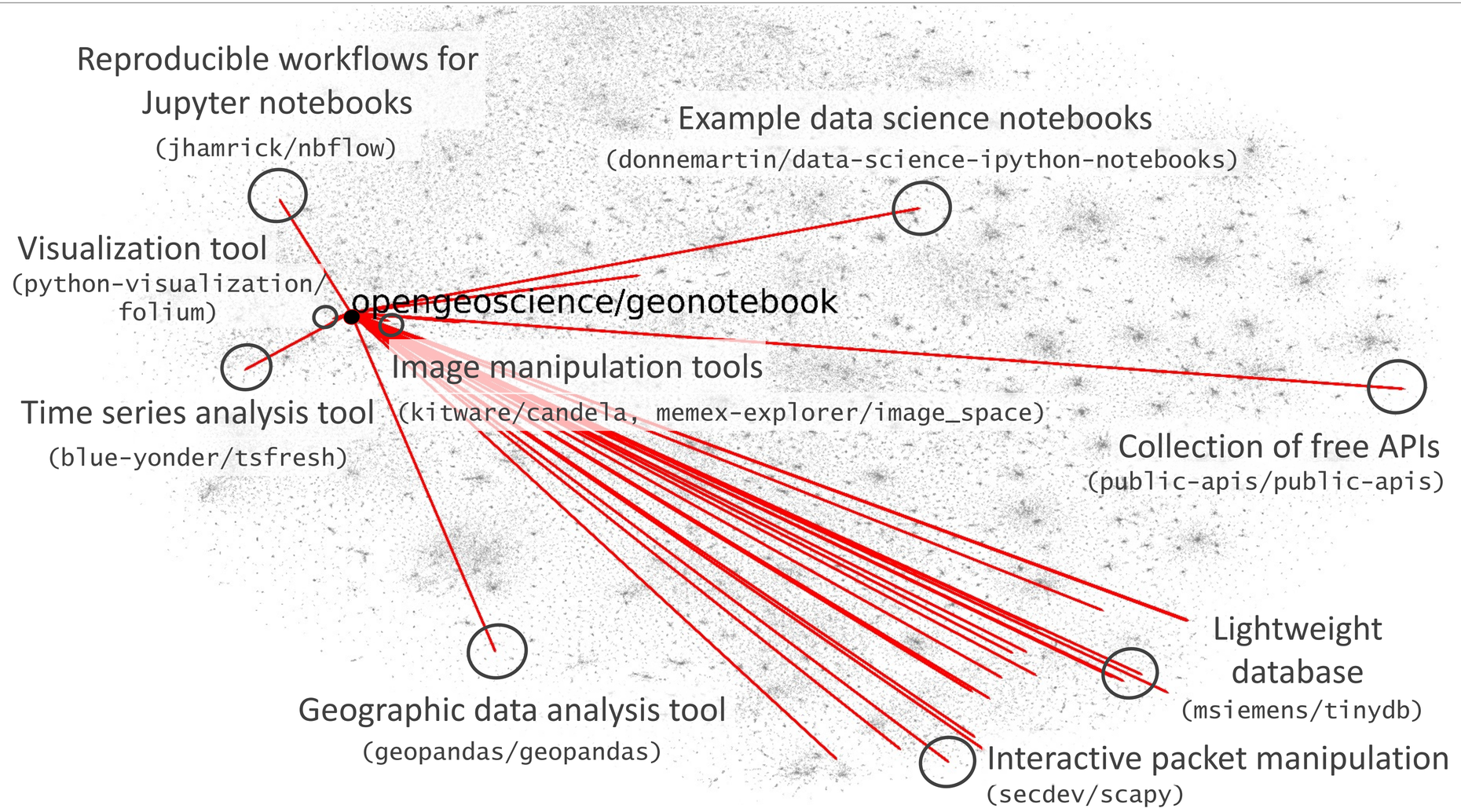


Do OSS developers also find their new ideas through weak ties?



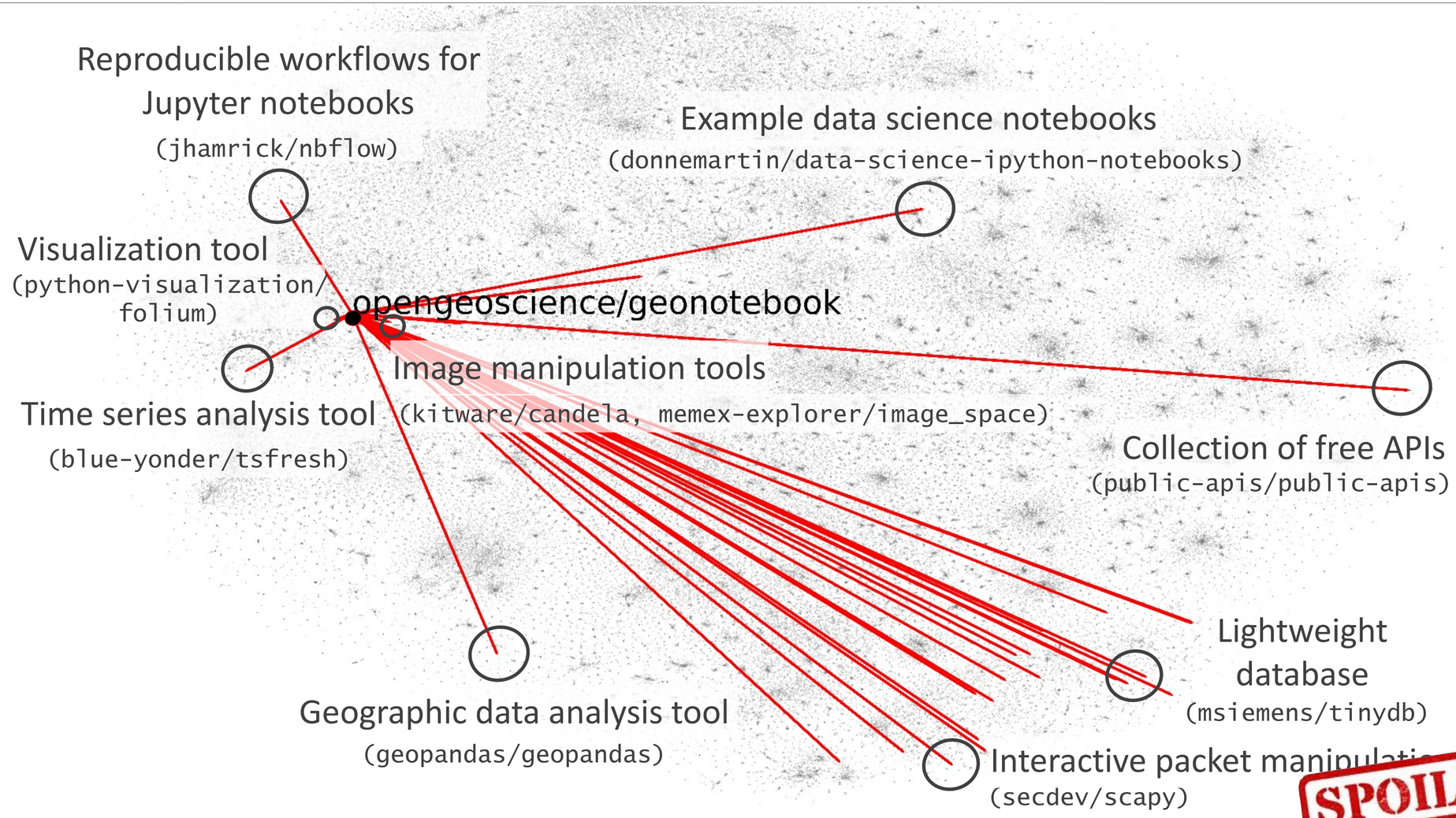
https://github.com/opengeoscience/geonotebook

Do OSS developers also find their new ideas through weak ties?



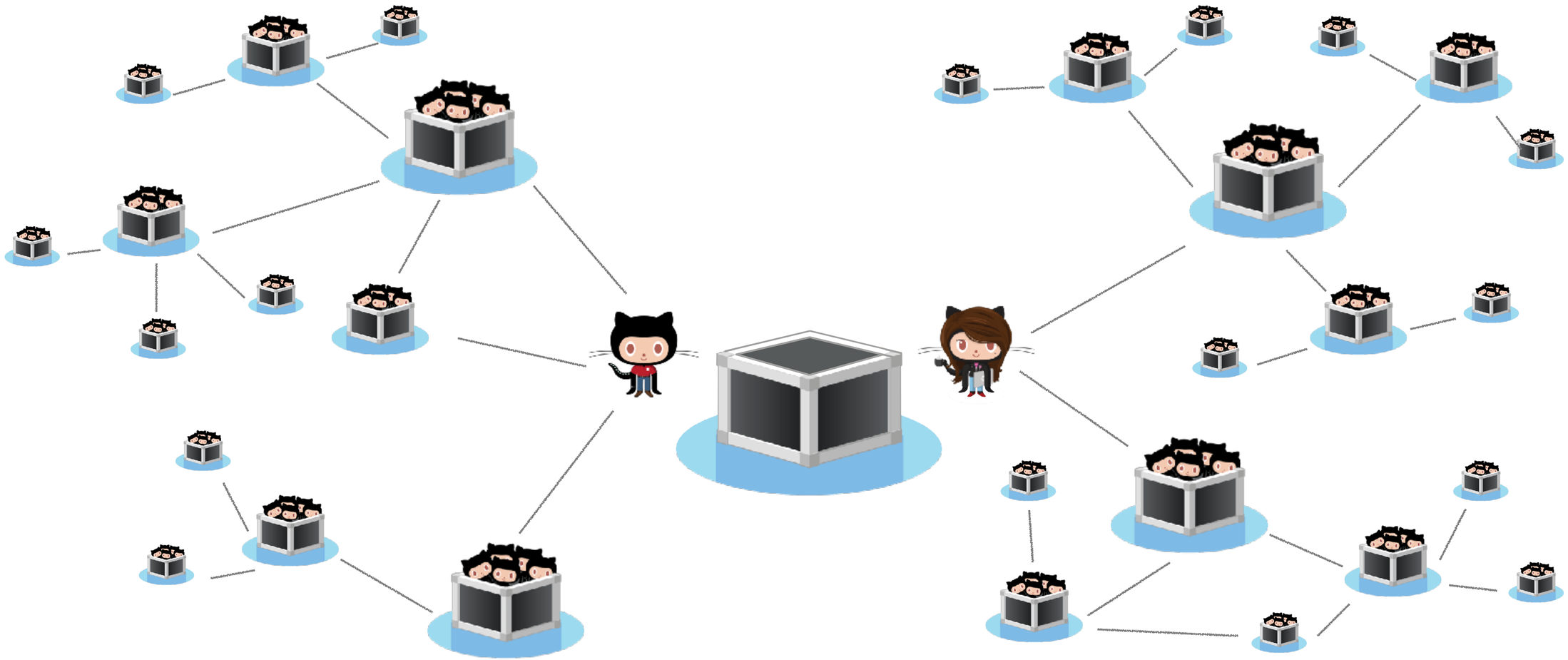
Anecdotally, yes

Do OSS developers also find their new ideas through weak ties?

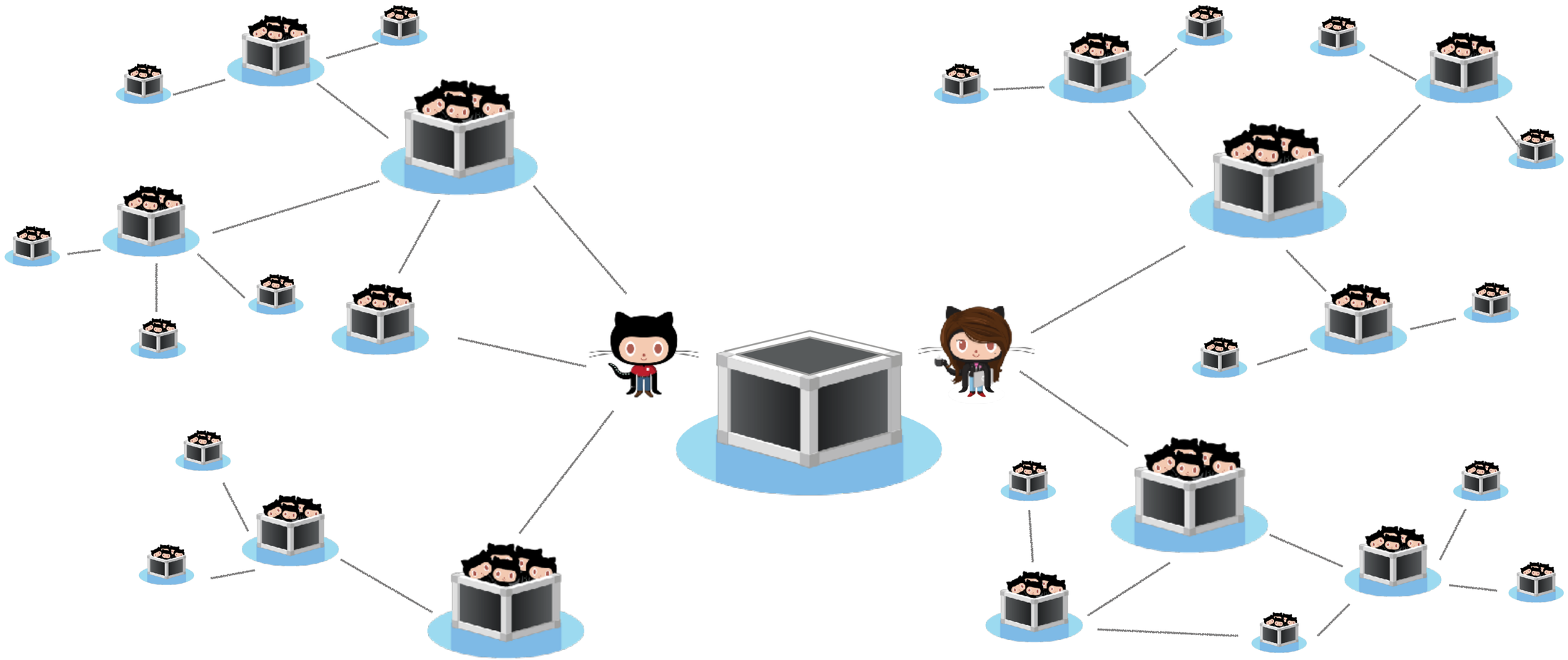


Amazingly, statistically also yes!

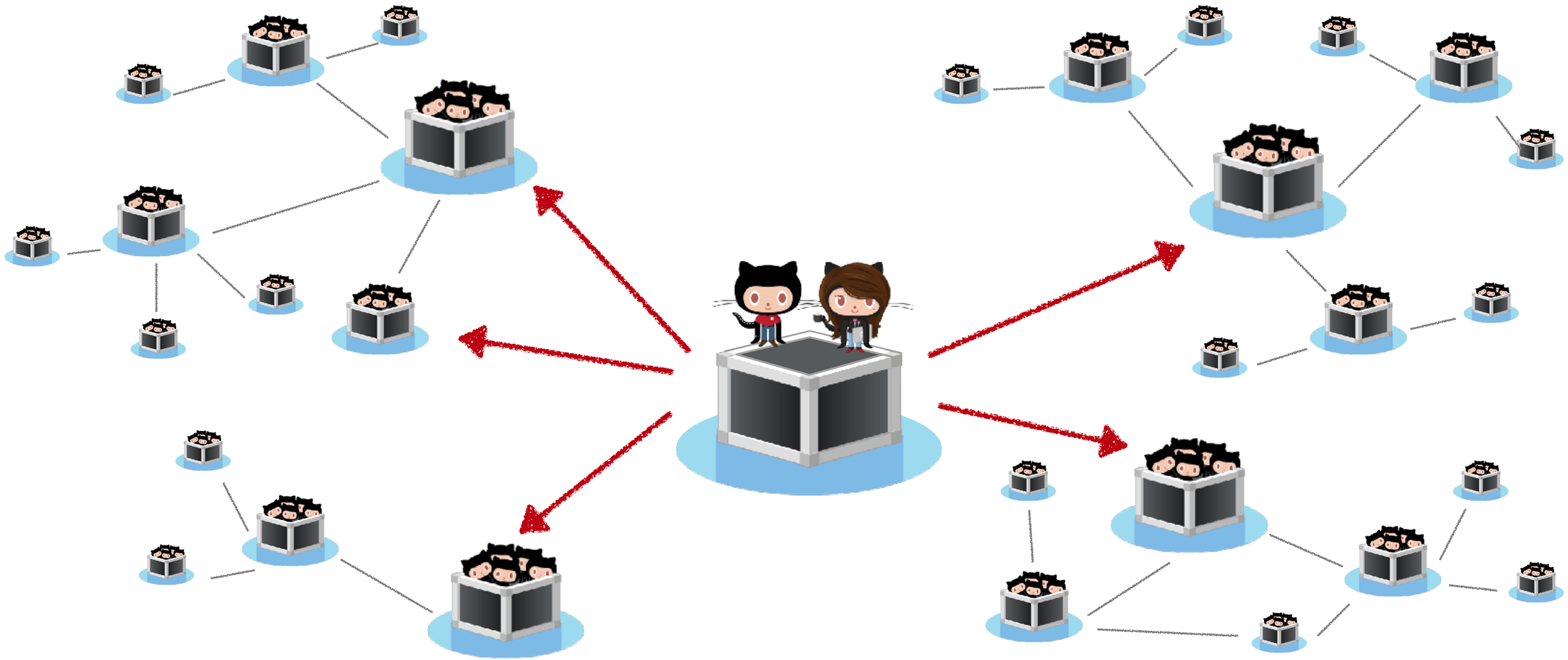
People interact with artifacts and with each other. This creates ties.



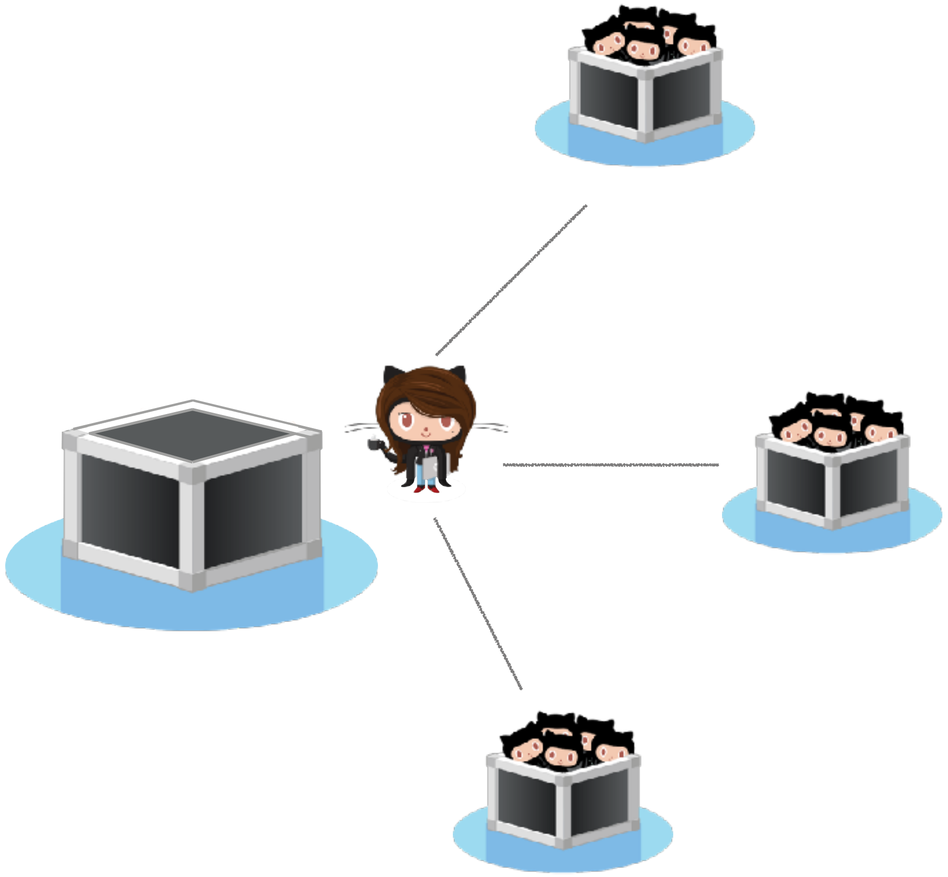
Hypothesis 1: The bigger developers' networks are, the better informed they are, and the more innovative their projects are.



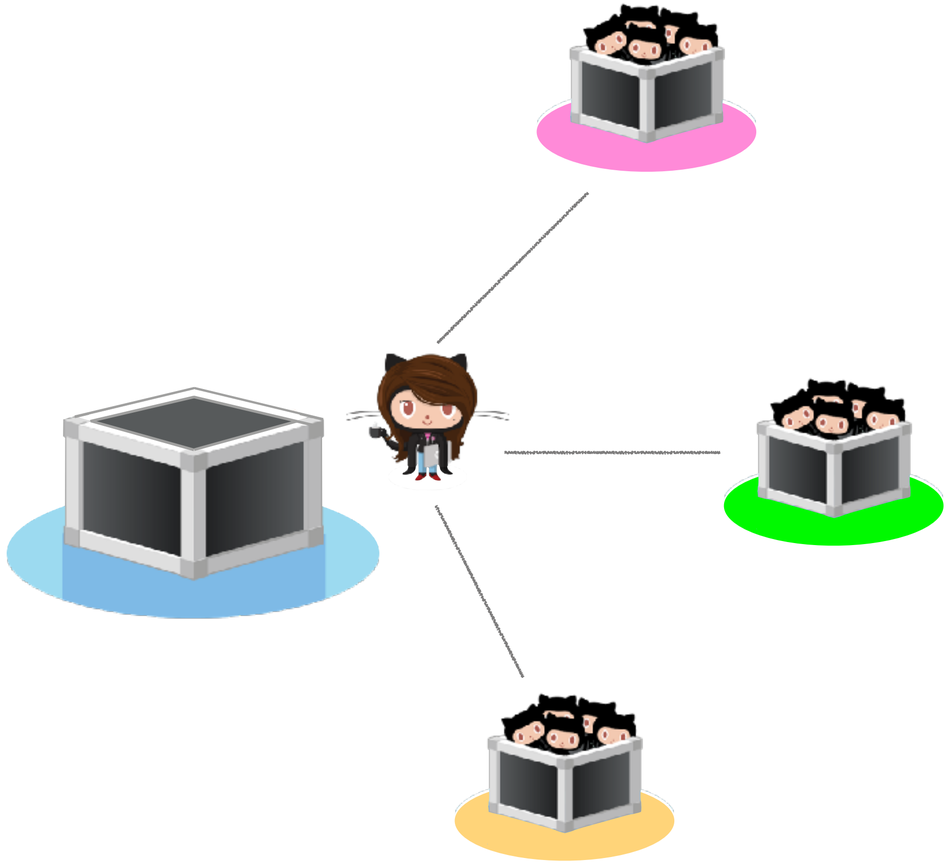
Measure: Out-degree centrality



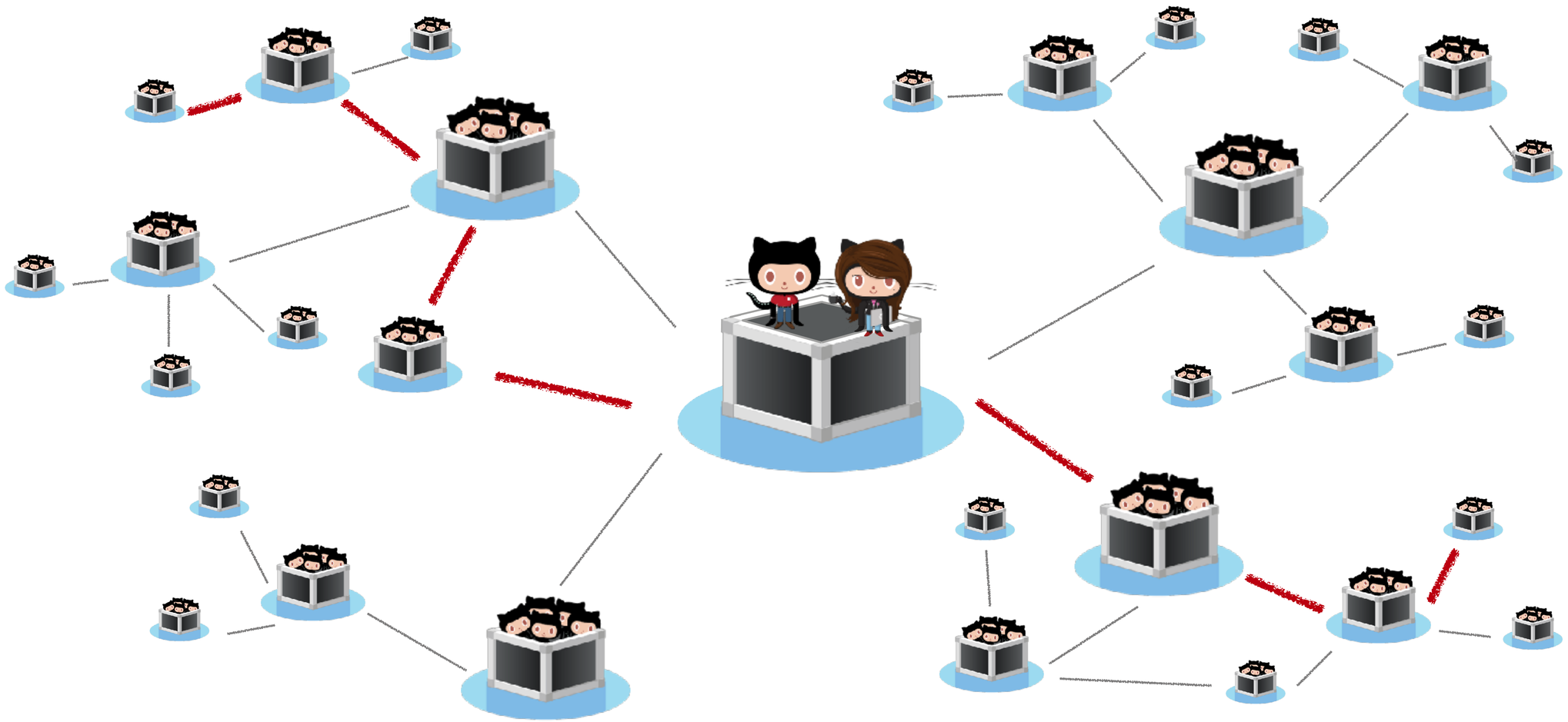
Hypothesis 2: The greater the informational diversity of developers' networks, the more innovative their projects are.



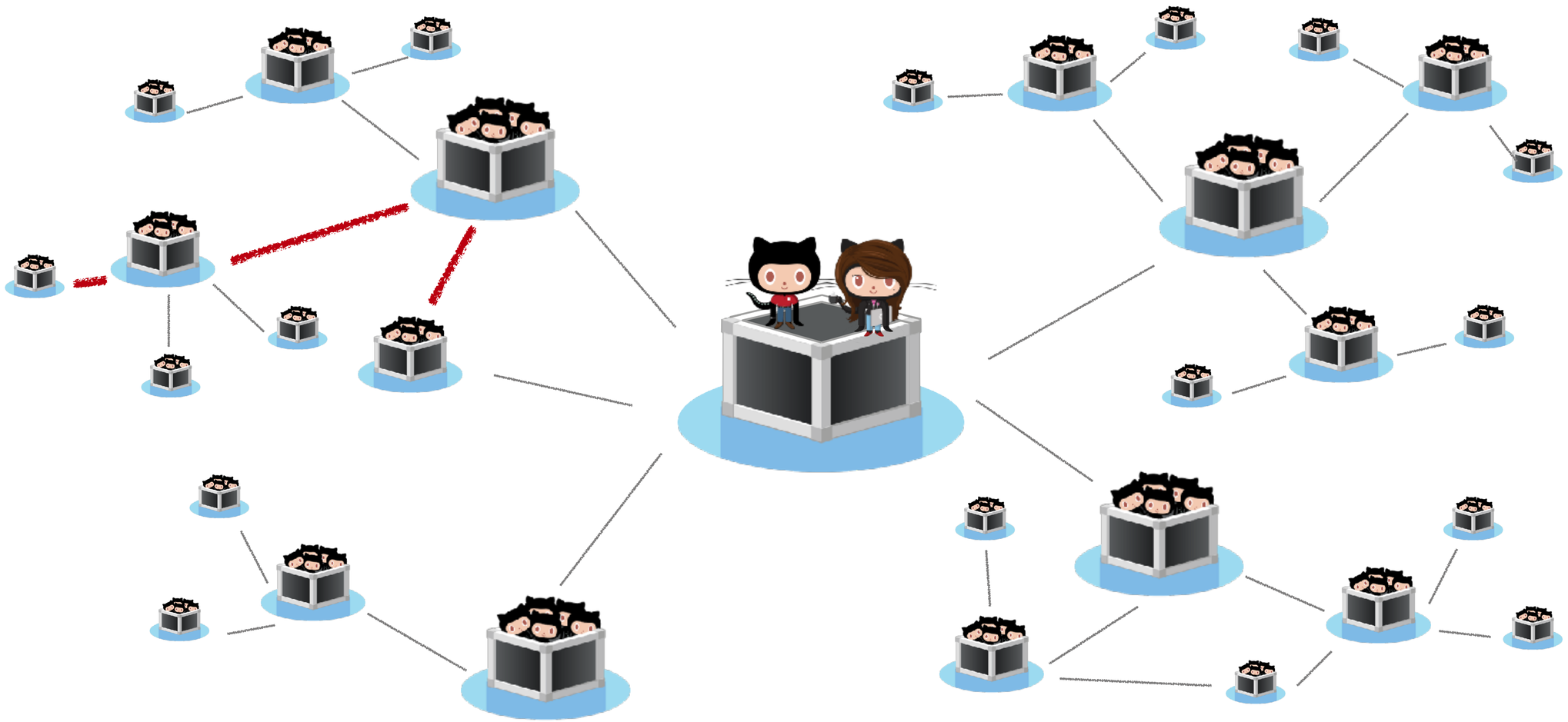
vs.



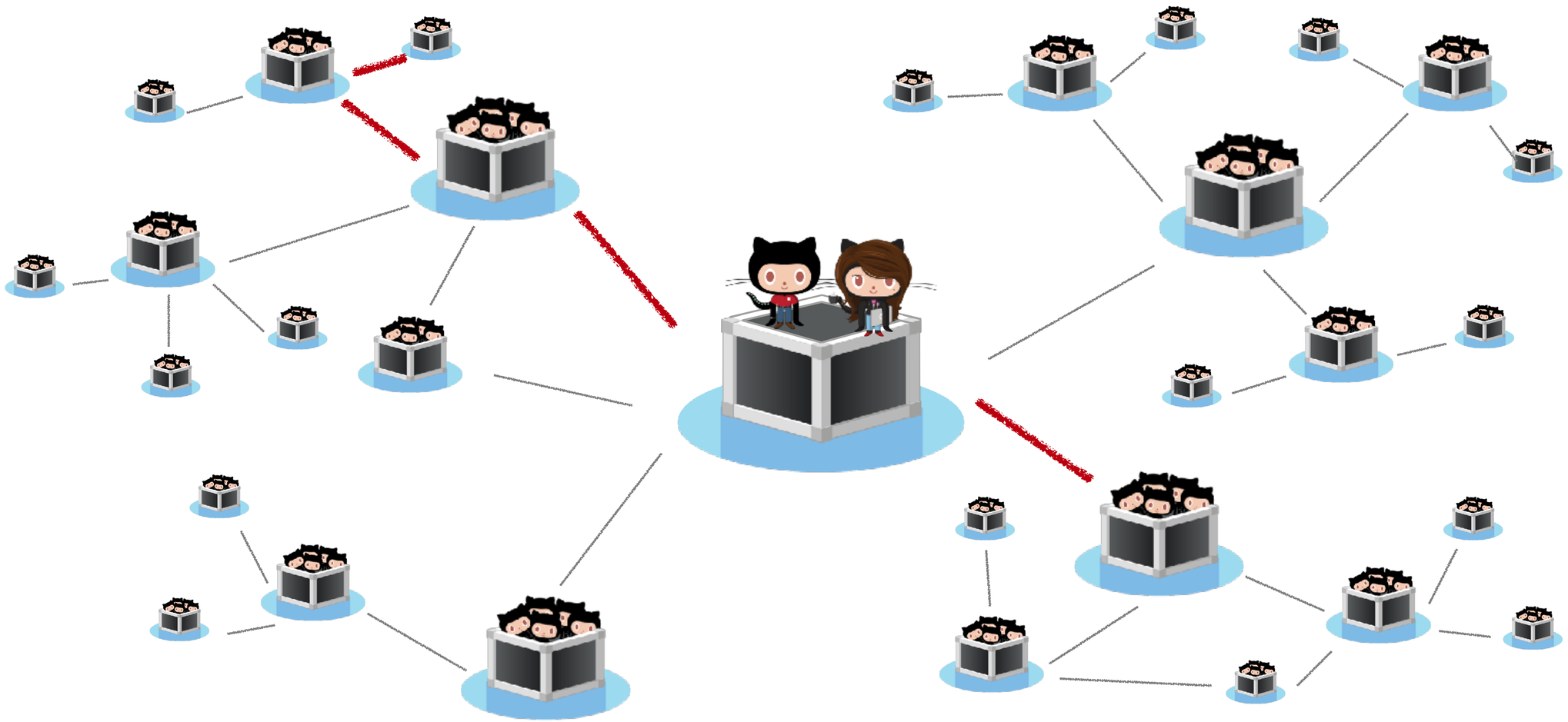
Measure: First, we generate Node2Vec embeddings for each project



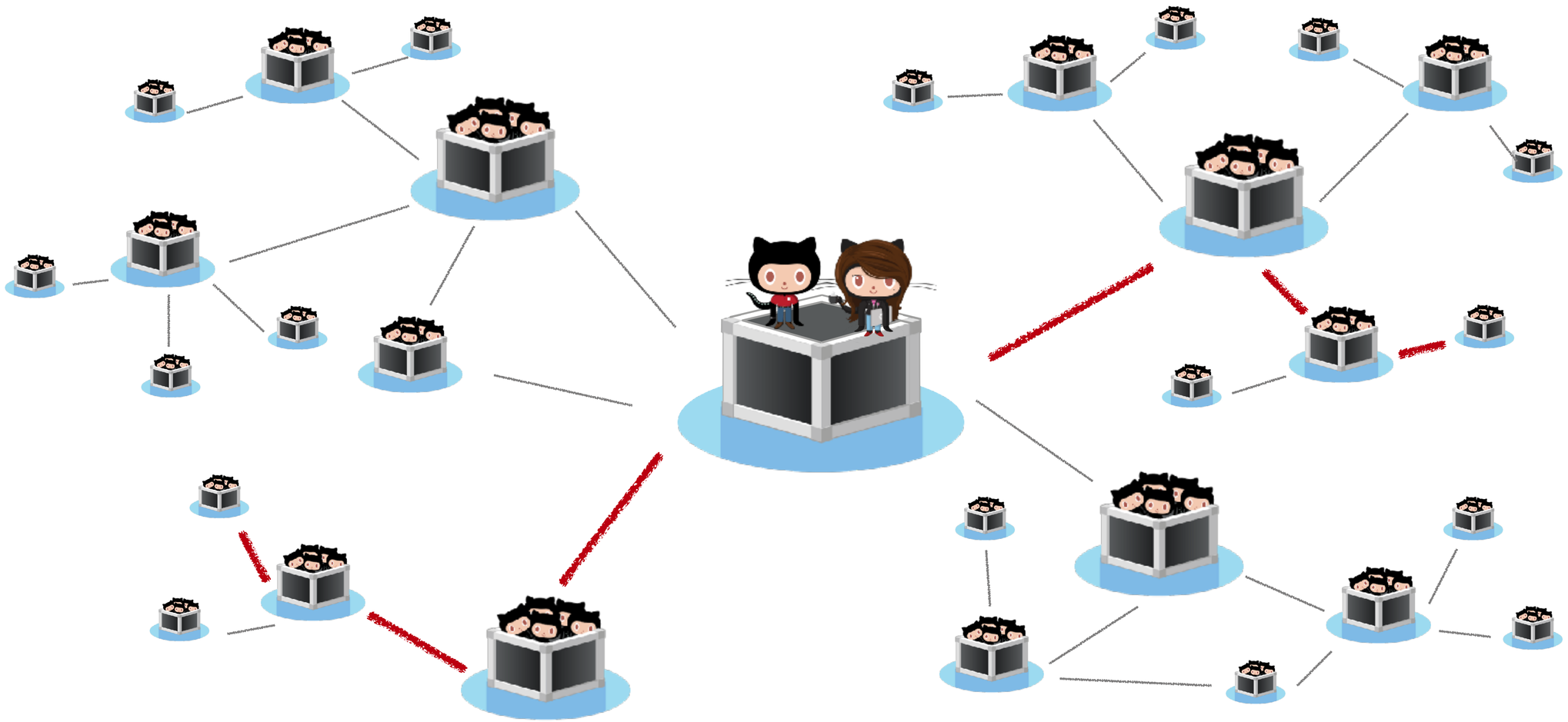
Measure: First, we generate Node2Vec embeddings for each project



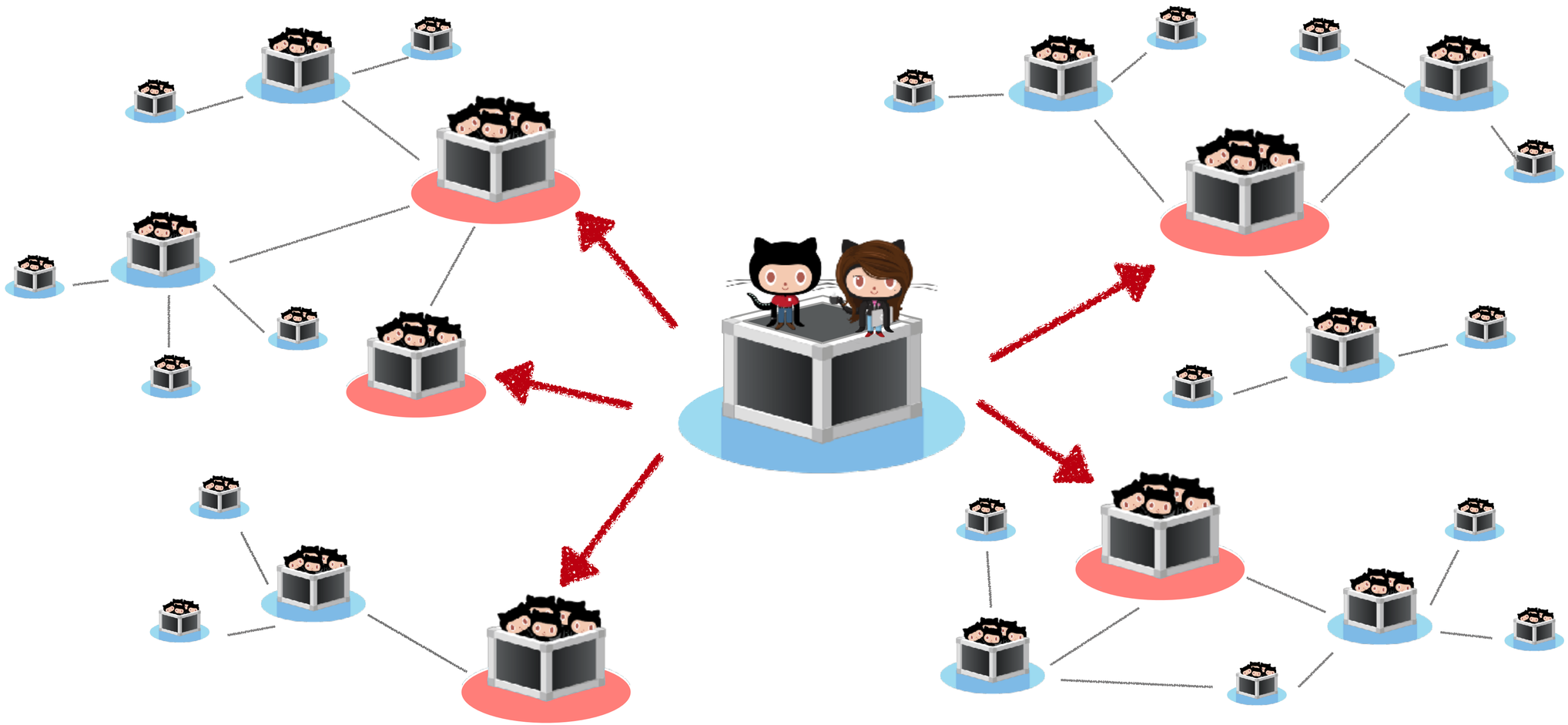
Measure: First, we generate Node2Vec embeddings for each project



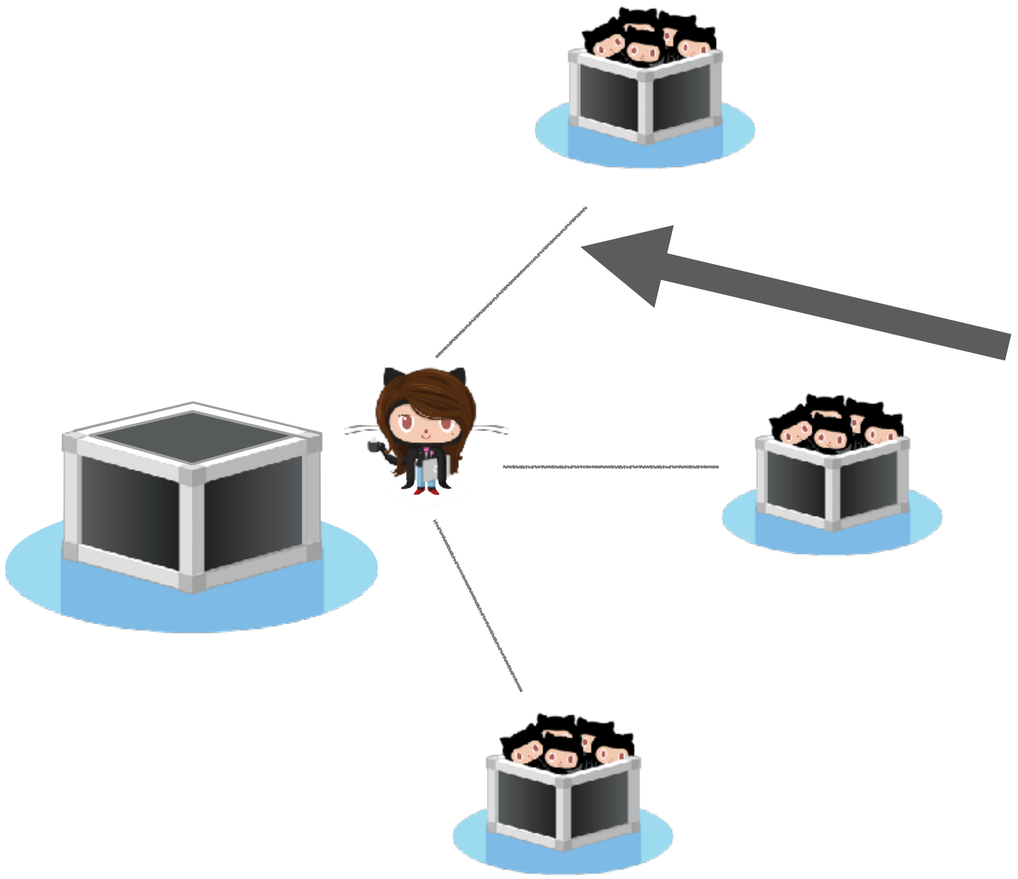
Measure: First, we generate Node2Vec embeddings for each project



Measure: Then, we compute the average pairwise distance (inverse cosine similarity) between a focal project's direct neighbors



From interactions to ties of varying strength



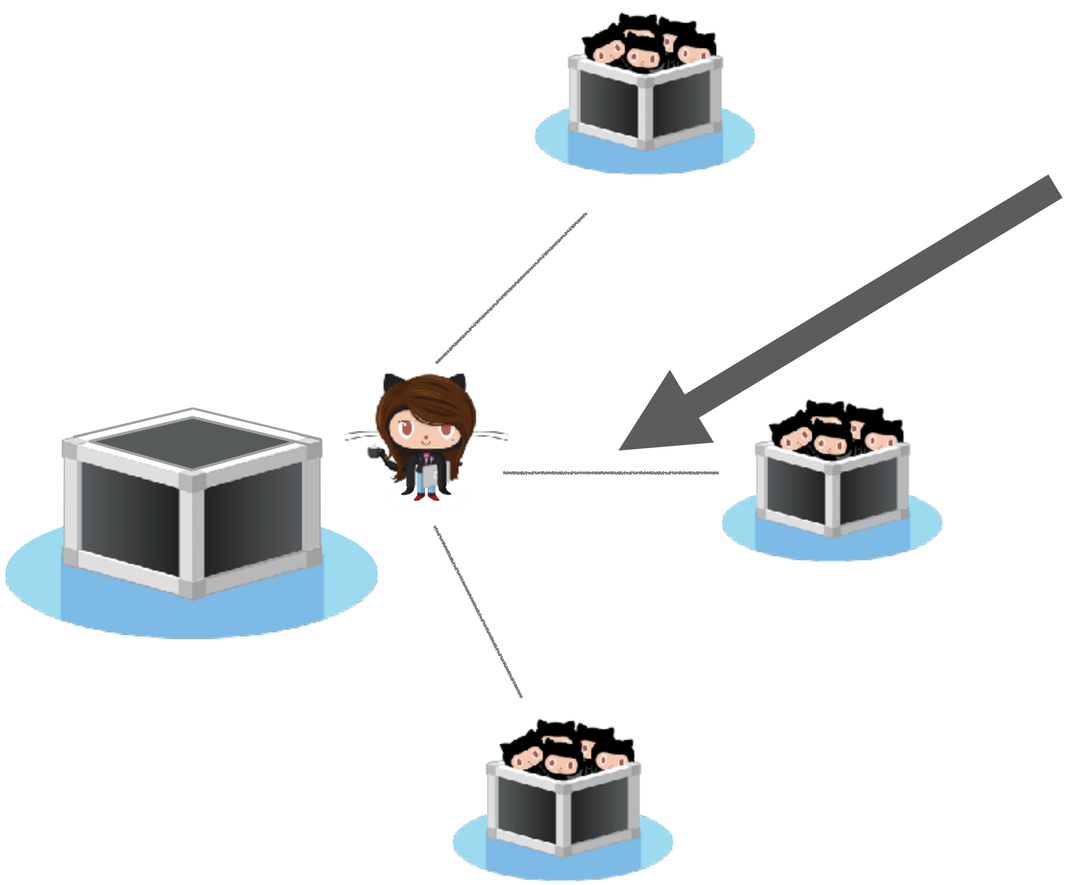
1 file changed +1 -1 lines changed

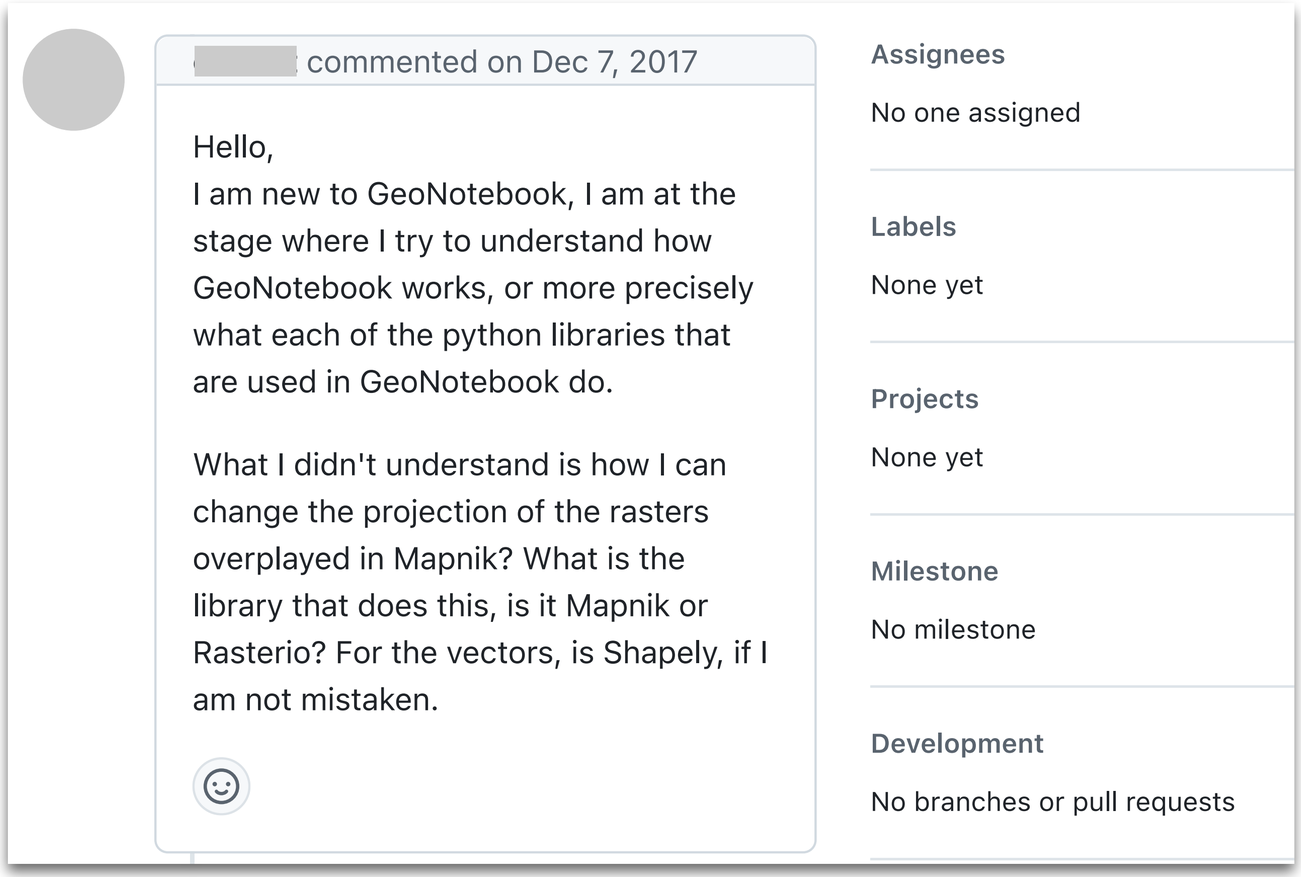
js/config/resolve.js +1 -1

```
@@ -1,6 +1,6 @@
1  var path = require('path');                    1  var path = require('path');
2                                                 2
3  - var renderer =                               3  + var renderer =
     process.env.GEONOTEBOOK_MAP_RENDERE            process.env.GEONOTEBOOK_MAP_RENDERE
     R || 'geojs';                                  R || 'ol';
4                                                 4
5  module.exports = {                             5  module.exports = {
6    alias: {                                     6    alias: {
```

Commits to the codebase
(relatively deep understanding of the codebase)

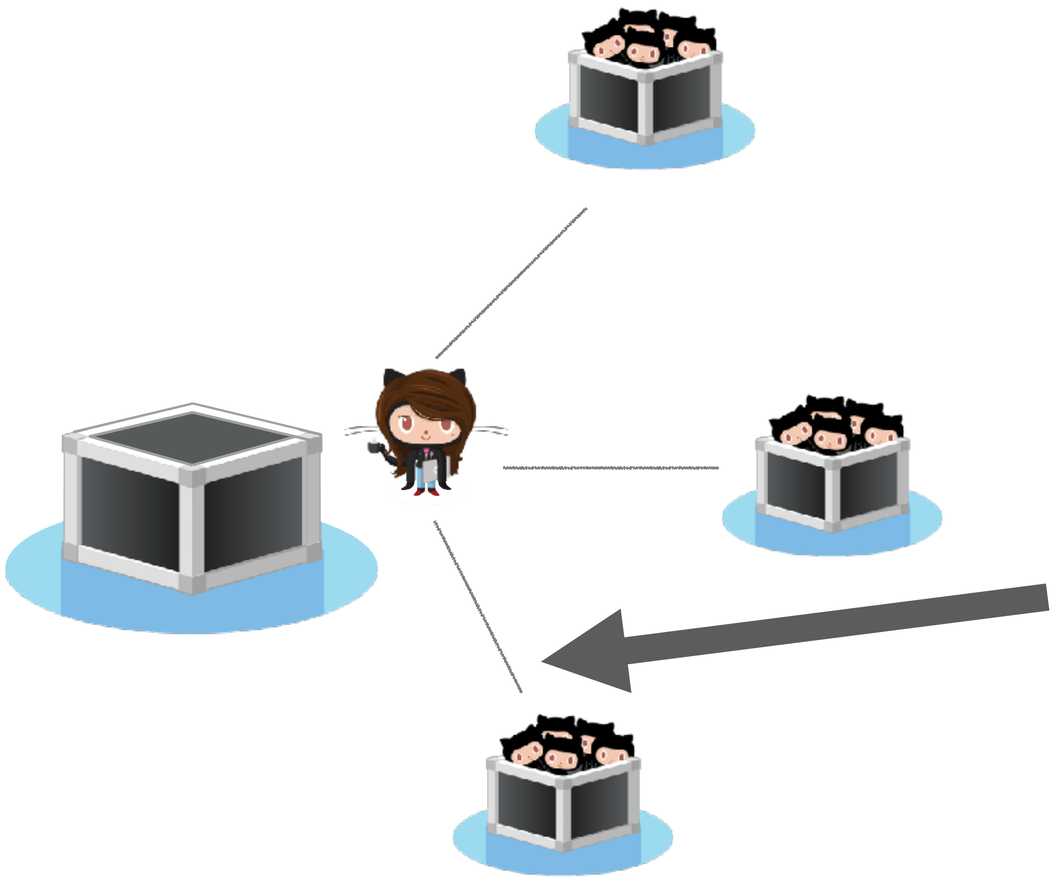
From interactions to ties of varying strength

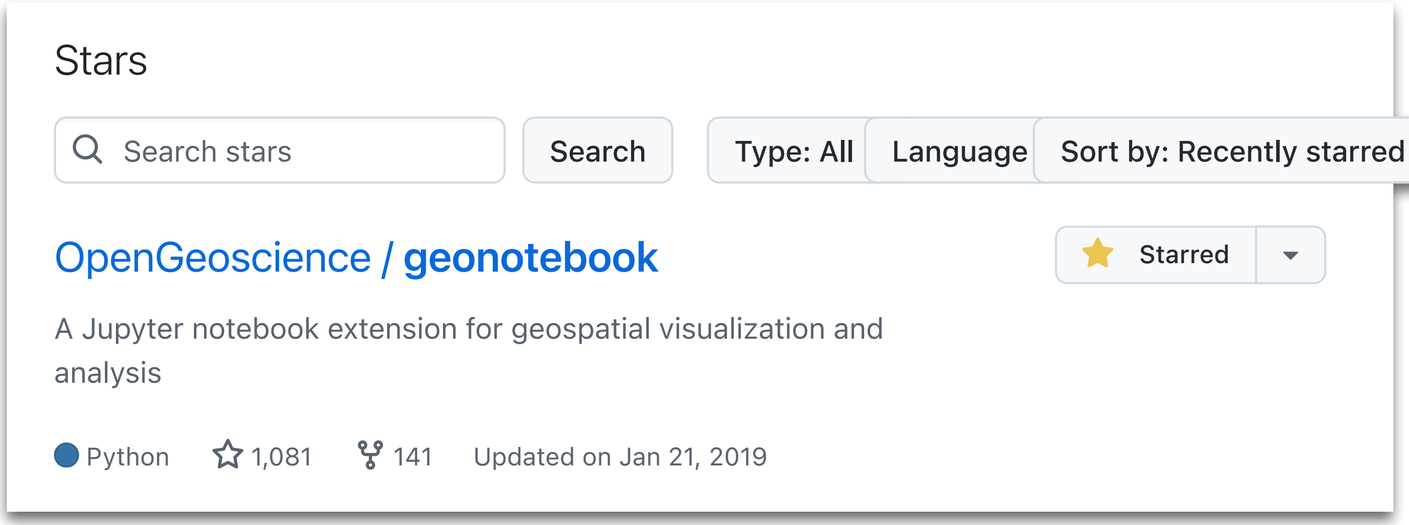




Issue reports
(some understanding of the project)

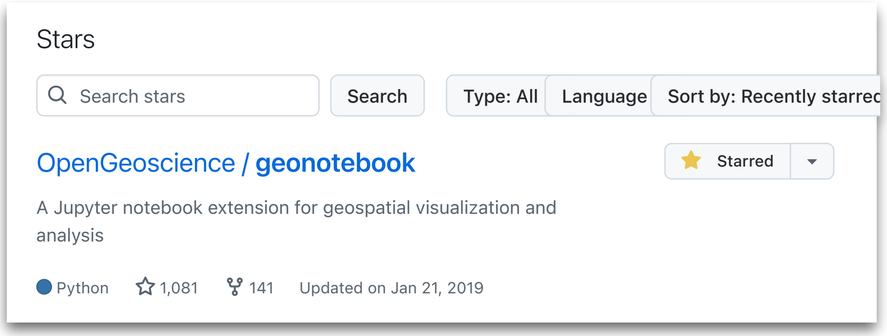
From interactions to ties of varying strength



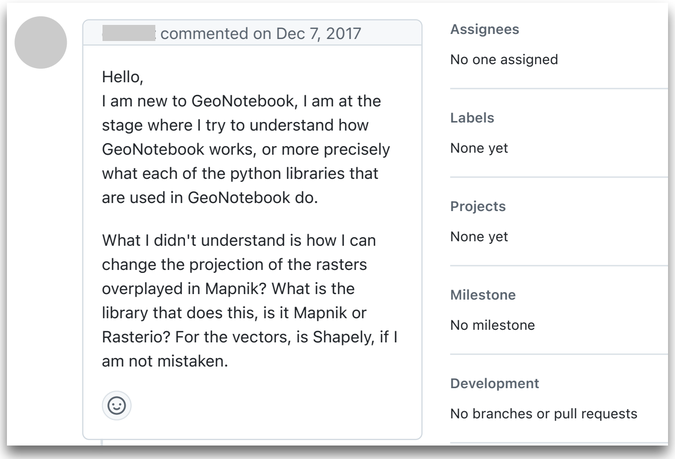


Stars
(awareness of the project)

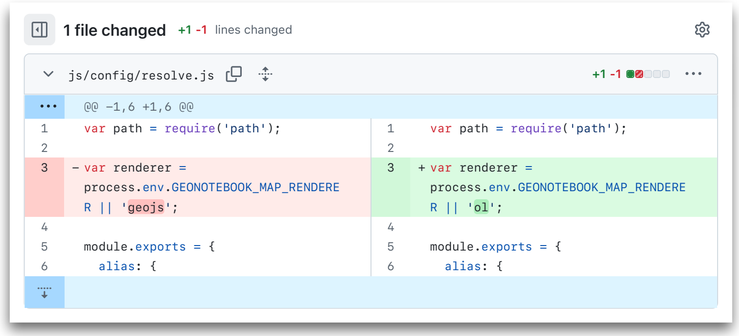
Many interactions are possible, these were just three examples.



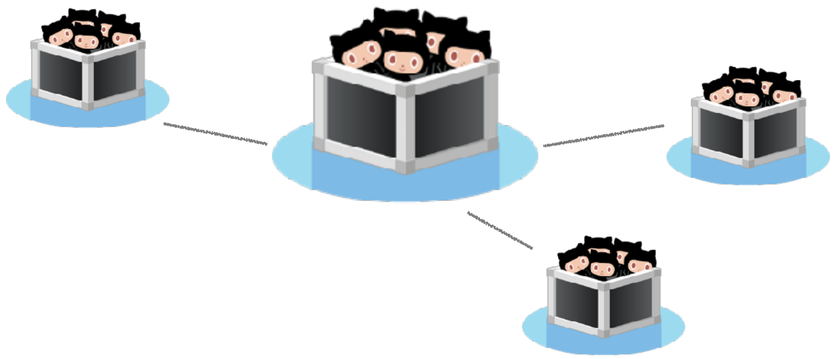
Stars



Issues



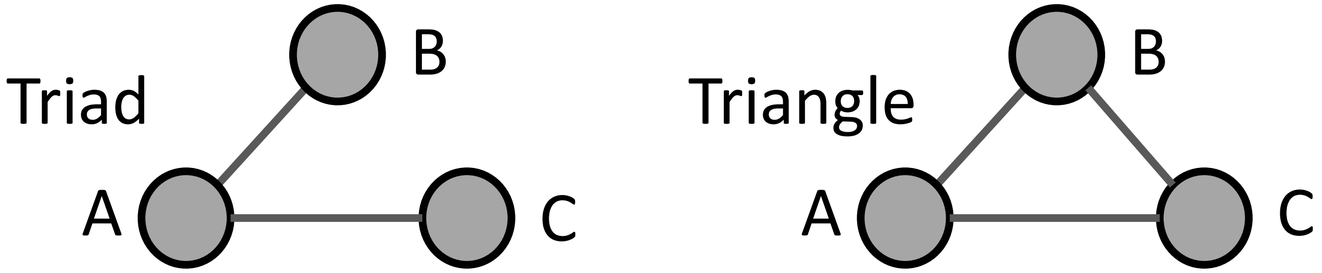
Commits



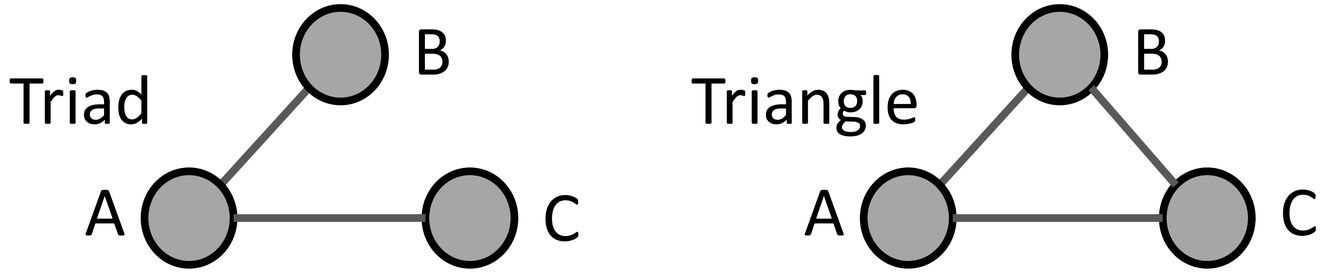
Weaker ties

Stronger ties

In strongly-tied social networks, triads are unlikely.



There is ~an order of magnitude (10×) difference in transitivity values between each pair of networks.

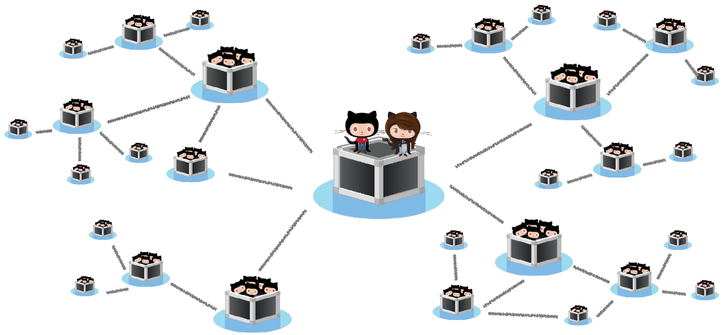


Interaction	**#Nodes**	**#Edges**	**Transitivity** $(\times 10^{-2})$
Commits	$763,062$	$1,926,978$	30.04
Issues	$278,945$	$727,255$	3.42
Stars	$480,394$	$3,658,543$	0.23

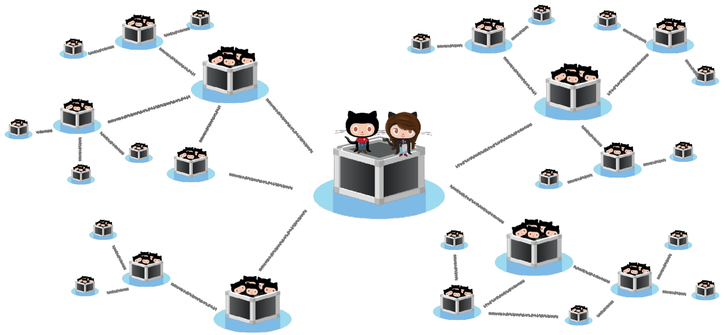
$$\text{Transitivity} = 3 * \text{N}_{\text{triangles}} / \text{N}_{\text{triads}}$$

Commits >> Issues >> Stars

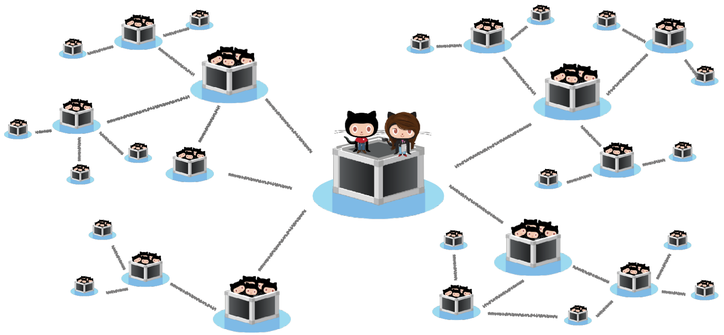
Now what?



Stars

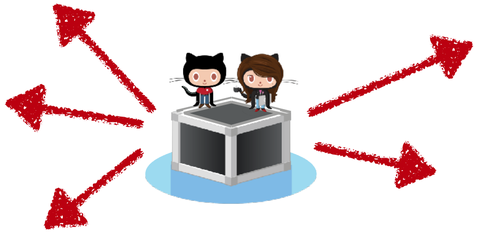


Issues

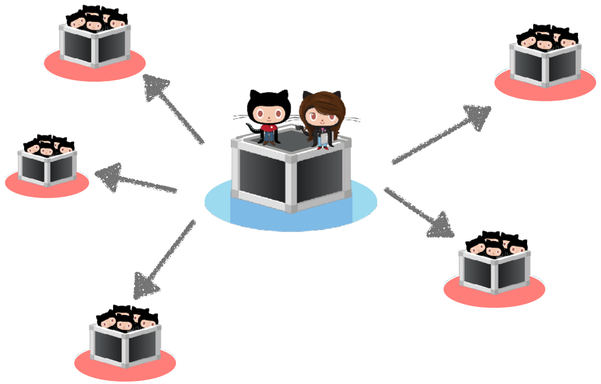


Commits

Out-degree centrality x 3?



Information diversity index x3?



The first two PCs cumulatively explain over 80% of the variance.

	Out-deg. centrality			Diversity index		
	PC1	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
D_{commit}	0.60	−0.45	0.67	0.63	−0.36	0.69
D_{issue}	0.61	−0.28	−0.74	0.64	−0.24	−0.72
D_{star}	0.52	0.85	0.11	0.43	0.90	0.08

PC1: Average volume of information available / Average diversity of the knowledge space (hyp 2)

The first two PCs cumulatively explain over 80% of the variance.

	Out-deg. centrality			Diversity index		
	PC1	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
D_{commit}	0.60	−0.45	0.67	0.63	−0.36	0.69
D_{issue}	0.61	−0.28	−0.74	0.64	−0.24	−0.72
D_{star}	0.52	0.85	0.11	0.43	0.90	0.08

PC2: Where the connectivity / diversity comes from

(The strength of weak ties)

Hypothesis 3: The more the informational diversity can be attributed to weak ties, the more innovative the projects are.

	Out-deg. centrality			Diversity index		
	PC1	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
D_{commit}	0.60	−0.45	0.67	0.63	−0.36	0.69
D_{issue}	0.61	−0.28	−0.74	0.64	−0.24	−0.72
D_{star}	0.52	0.85	0.11	0.43	0.90	0.08

PC2: Where the connectivity / diversity comes from

(The strength of weak ties)

Finally, the novelty regression:

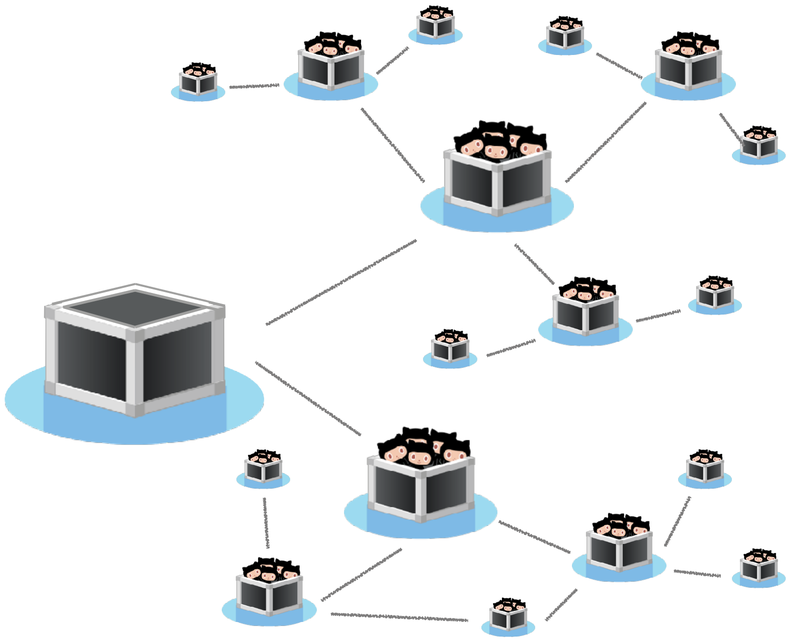
- Hypothesis 1 (greater connectivity): weak/inconsistent effects
- Hypothesis 2 (greater info diversity): small but clear effects (25—75 percentile: 4% change in the distribution)
- Hypothesis 3 (strength of weak ties): clear effects, comparable size

	Model III	Model IV
Variables of interest		
Deg_{ave} $(\mathbf{H}_1)$		-0.002^{***}
		(0.001)
$Deg_{weakness}$		-0.005^{***}
		(0.001)
Div_{ave} $(\mathbf{H}_2)$	0.007^{***}	0.008^{***}
	(0.001)	(0.001)
$Div_{weakness}$ $(\mathbf{H}_3)$	0.005^{***}	0.007^{***}
	(0.001)	(0.001)
Observations	38,164	38,164

$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Exposure to diverse ideas through weak ties predicts novel combinations of packages.

- Lurking on the GitHub platform seems to have quantifiable benefits. Redesign the Trending page?
- Automated project recommendation tools may be counterproductive?
- Well-informed but not necessarily highly active developers may also be experts at their craft?
- How to track and give credit to ideas?
- Surface-level vs deep-level diversity?
- AI-generated code: novel or regression to the mean?



Today: Let’s look at some concrete examples of network effects

Measuring innovation in software

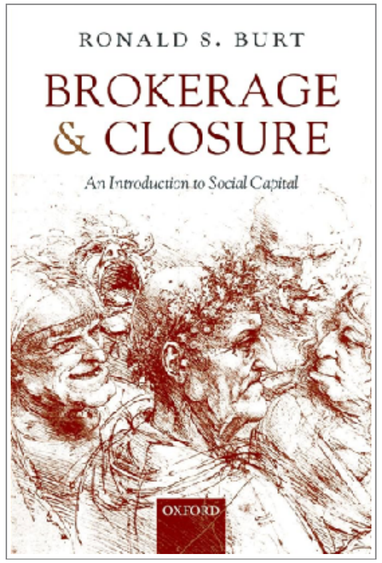
Understanding how innovation emerges

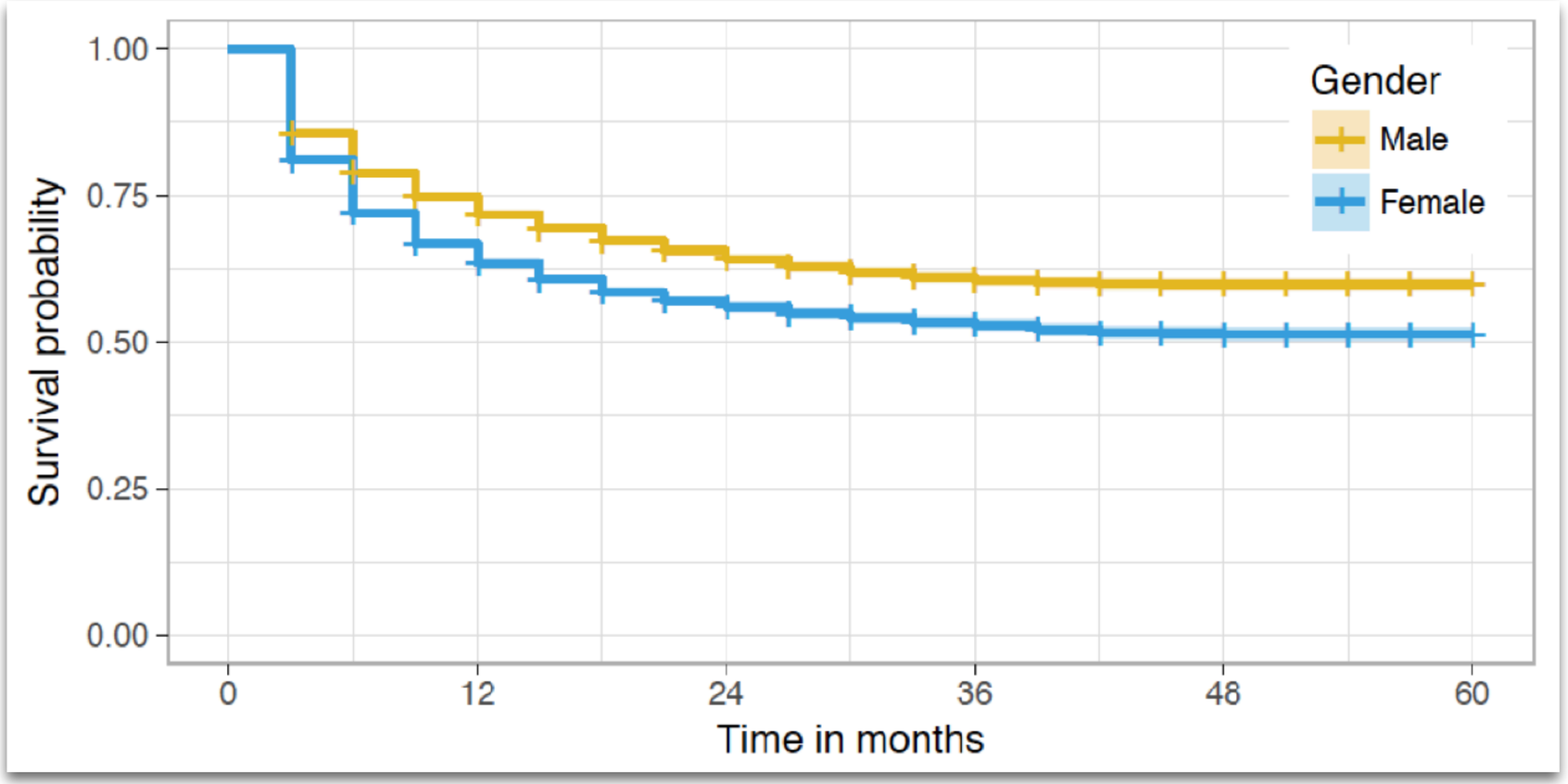
Social capital

Social contagion

Weak ties predict longer-term participation for women

Social capital mechanism





Qiu, Nolte, Brown, Serebrenik, and Vasilescu. “Going farther together: The impact of social capital on sustained participation in open source.” ICSE 2019 Distinguished Paper Award.

Weak ties predict the spread of tools

Diffusion of innovations mechanism

12 popular quality assurance tools

Continuous integration	*Dependency management*	*Code coverage reporters*	*Cross browser testers*
build passing	dependencies up to date	coverage 94%	Firefox 82 ✓ Chrome 86 ✓
Travis	David	Coveralls	Saucelabs
Circle	Bithound	Codeclimate	
Appveyor	Gemnasium	Codecov	
Codeship		Codacy	

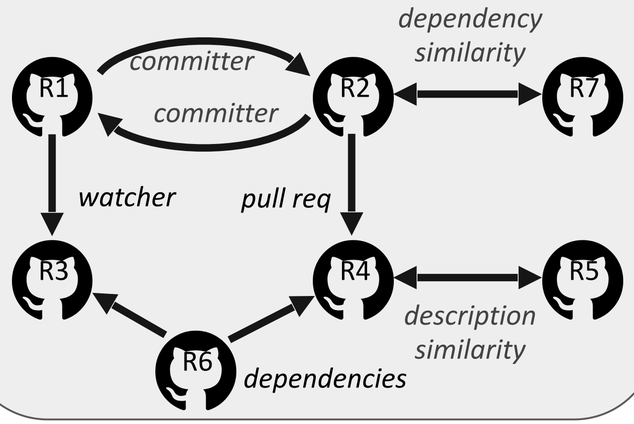
~86,000 npm package repositories

GitHub

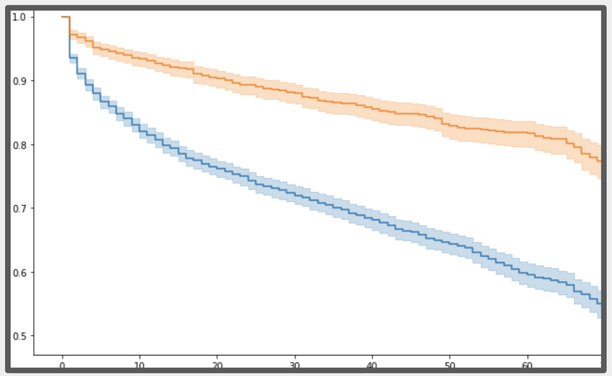
npm

For each tool:

Heterogeneous network



Hazard modeling (Cox regression)



Lamba, Trockman, Armanios, Kästner, Miller, and Vasilescu. “Heard it through the Gitvine: An empirical study of tool diffusion across the npm ecosystem. FSE 2020.

Acknowledgements


Courtney Miller


Anita Brown


Asher Trockman


Jim Herbsleb


Shurui Zhou

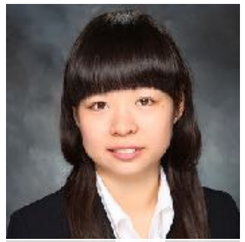

Hongbo Fang


Anita Sarma


Cassandra Overney


Audris Mockus


Alex Nolte


Sophie Qiu

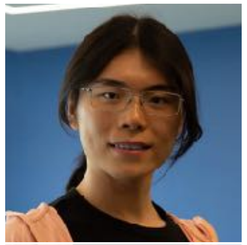

Alex Serebrenik


Marat Valiev


Laura Dabbish


Lily Li


Naveen Raman


Hao He


Christian Kästner


Hemank Lamba


Emerson Murphy-Hill

Google

Alfred P. Sloan FOUNDATION

National Science Foundation

FORDFOUNDATION

STRUDEL sustainability research on …

Project practices

- CHASE 2023 (social media)
- ICSE 2020 (forking)
- ESEC/FSE 2019 (forking)
- ESEC/FSE 2018 (abandonment factors)

Attracting contributors

- ICSE 2022 (Twitter)
- MSR 2020 (Twitter)
- CSCW 2019 (signals)
- ESEC/FSE 2015 (social connections)

Stress, burnout, disengagement

- ICSE 2022 (toxicity theory)
- ICSE SEIS 2022 (toxicity vs pushback)
- ICSE NIER 2020 (toxic language)
- ICSE 2019 (overwork)
- OSS 2019 (dropout, survival analysis)

Novelty and innovation

- ICSE 2024 (atypical combinations)

Funding models

- ICSE 2020 (donations)

Transparency and signaling

- ESEC/FSE 2020 (diffusion of practices)
- CSCW 2019 (signals)
- ICSE 2018 (badges)

Diversity and inclusion

- CHI 2023 (ClimateCoach)
- ICSE SEIS 2023 (census)
- ICSE 2019 (social capital)
- CHI 2015 (gender & tenure)
- CHASE 2015 (survey)

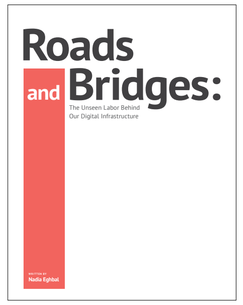
Network effects

- ICSE 2024 (innovation)
- ESEC/FSE 2023 (labor pools)
- ICSE 2022 (Twitter)
- ESEC/FSE 2020 (diffusion of practices)
- ICSE 2019 (social capital)
- ESEC/FSE 2018 (abandonment factors)

Sunsetting

- ESEC/FSE 2023 (dealing with abandonment)

Like any infrastructure, it needs regular upkeep and maintenance

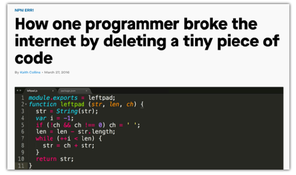


Everybody uses open source:

- Fortune 500 companies
- Major software companies
- Startups
- Government
- …

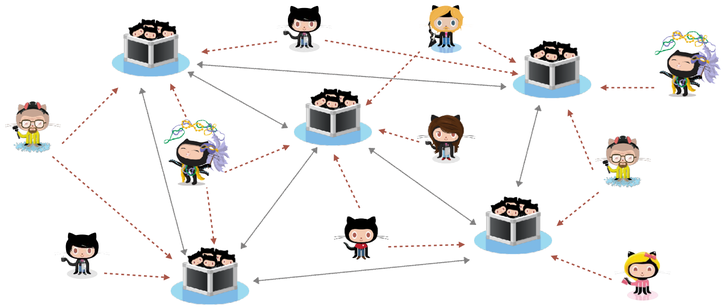
If undermaintained:

- Brittle supply chains
- Risks for downstream users
- Slows down innovation
- …



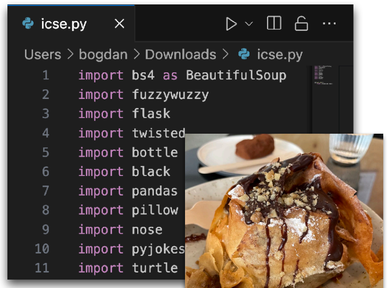
https://qz.com/646467/how-one-programmer-broke-the-internet-by-deleting-a-tiny-piece-of-code/

Contributors and projects form complex socio-technical networks!



Can we measure the network effects?

Software innovation as novel recombination of software libraries



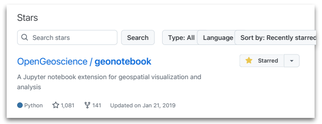
Lots of combinations:

- **(twisted, bottle)**
- **(turtle, nose)**
- **(black, pandas)**
- **(fuzzywuzzy, pillow)**
- …

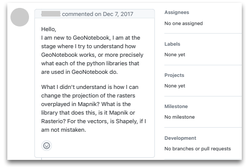
C(n,2) unique pairs of packages.

Dark chocolate + apple strudel is arguably innovative because it is atypical.

Exposure to diverse ideas through weak ties predicts novel combinations of packages.

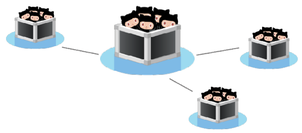


Stars



Issues

Commits



Weaker ties

Stronger ties

DALL·E 3 - “Networks of open-source software projects”

The Strength of Weak Ties in Open-Source Software Development Networks

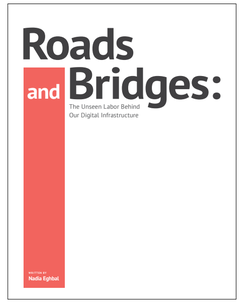
Bogdan Vasilescu
At UC Irvine, October 3rd, 2024

STRUDEL
SOCIO-TECHNICAL RESEARCH USING DATA EXCAVATION LAB

Carnegie Mellon University

S3D
Software and Societal Systems Department

Like any infrastructure, it needs regular upkeep and maintenance

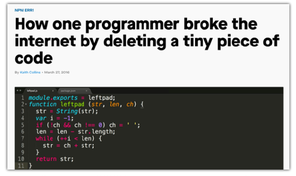


Everybody uses open source:

- Fortune 500 companies
- Major software companies
- Startups
- Government
- …

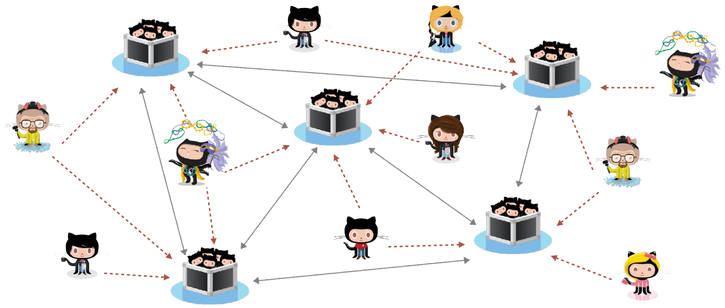
If undermaintained:

- Brittle supply chains
- Risks for downstream users
- Slows down innovation
- …



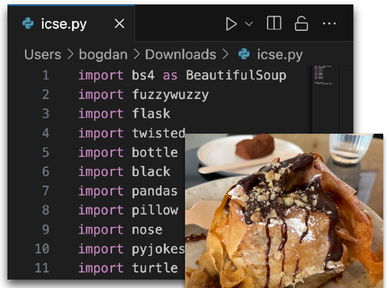
https://qz.com/646467/how-one-programmer-broke-the-internet-by-deleting-a-tiny-piece-of-code/

Contributors and projects form complex socio-technical networks!



Can we measure the network effects?

Software innovation as novel recombination of software libraries



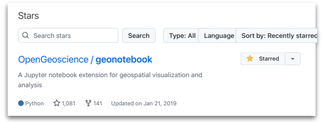
Lots of combinations:

- **(twisted, bottle)**
- **(turtle, nose)**
- **(black, pandas)**
- **(fuzzywuzzy, pillow)**
- …

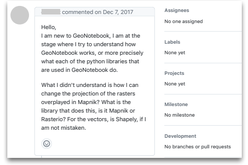
C(n,2) unique pairs of packages.

Dark chocolate + apple strudel is arguably innovative because it is atypical.

Exposure to diverse ideas through weak ties predicts novel combinations of packages.

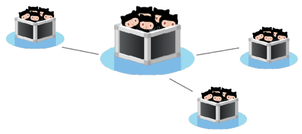


Stars



Issues

Commits



Weaker ties

Stronger ties

DALL·E 3 - “Networks of open-source software projects”

The Strength of Weak Ties in Open-Source Software Development Networks

Bogdan Vasilescu
At UC Irvine, October 3rd, 2024

STRUDEL
SOCIO-TECHNICAL RESEARCH USING DATA EXCAVATION LAB

Carnegie Mellon University

S3D
Software and Societal Systems Department