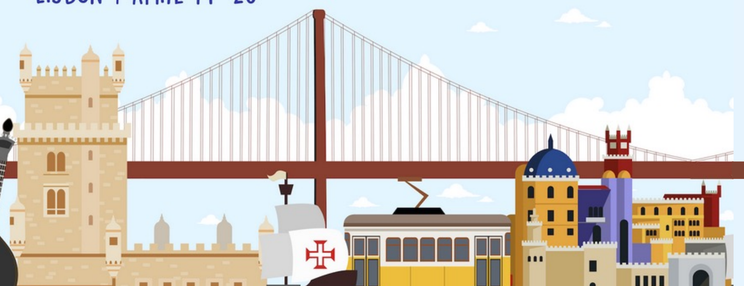
Novelty Begets Long-Term Popularity, But Curbs Participation

Lisbon, April 19, 2024

Hongbo Fang
@fang_hongbo

Jim Herbsleb
@jherbsleb

Bogdan Vasilescu
@b_vasilescu



STRUDEL

S3D Software and Societal Systems Department

Carnegie Mellon University

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”

How to define innovation in software?

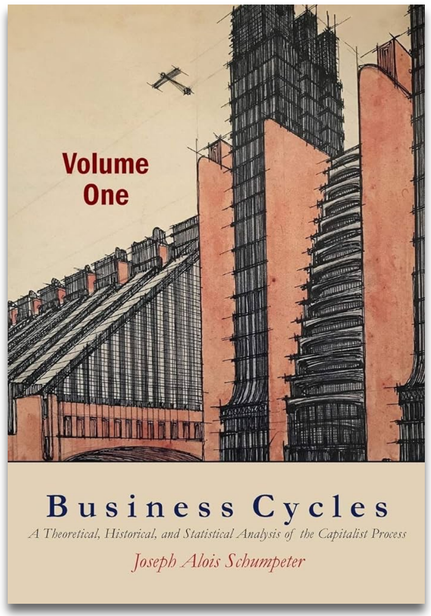
How to measure it?

How does innovation emerge?

What are its consequences?

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

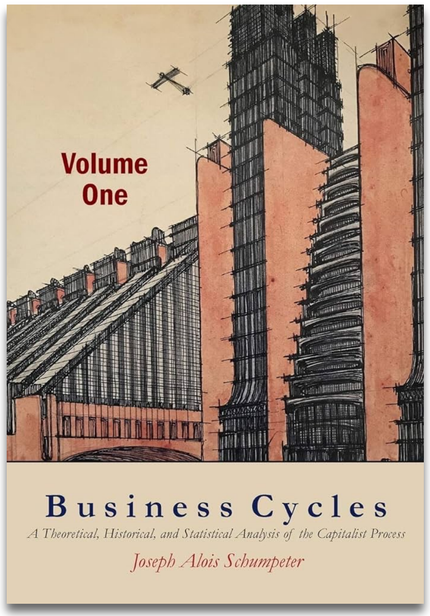
Key idea: Innovation as novel recombination



(Schumpeter, 1939)

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

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

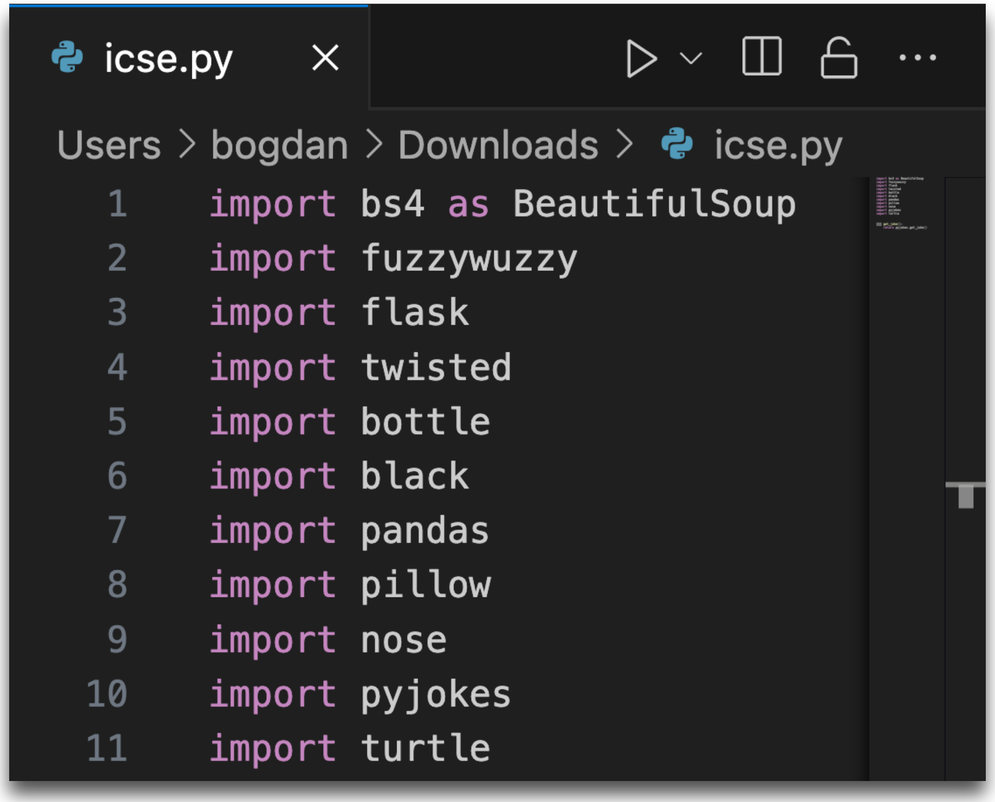
[1]Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, USA. [2]Northwestern Institute on Complex Systems, Northwestern University, 600 Foster, Evanston, IL 60208, USA. [3]Datascope Analytics, 180 West Adams Street, Chicago, IL 60603, USA. [4]National Bureau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

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

(Uzzi et al, 2013)

The naturalness line of work is a novel recombination of ideas from linguistics, NLP, software engineering, …

Software innovation as novel recombination of software libraries



A project importing n packages has C(n,2) package combinations:

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

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

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/

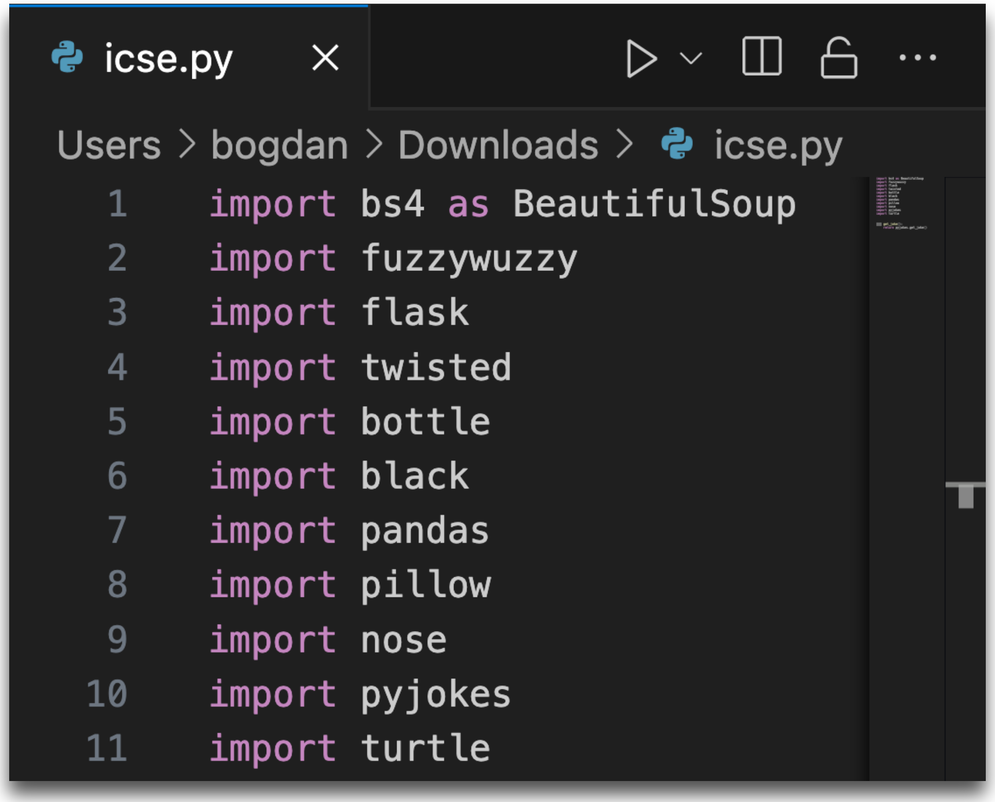
Software innovation as novel recombination of software libraries

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.
  - People may be impressed by your culinary creativity.
- Hyp: More innovative projects are less sustainable.
  - 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/

But how to measure the (a)typicality of a package combination?



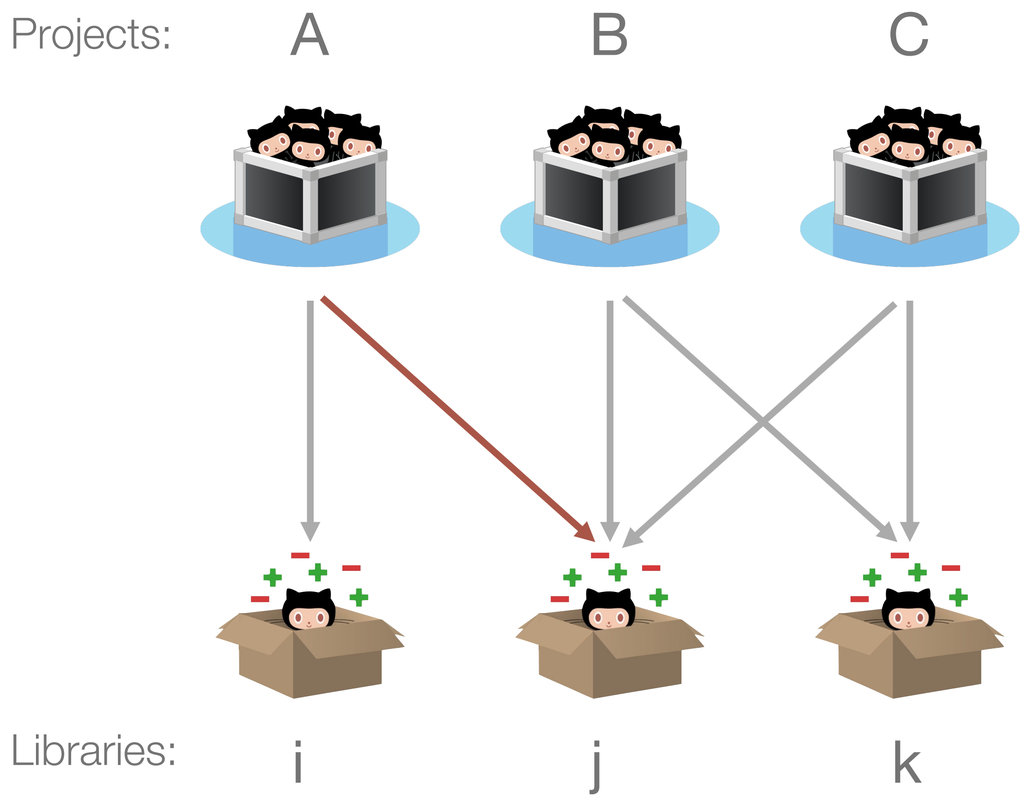
A project importing n packages has C(n,2) package combinations:

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

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

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

Observed reality:

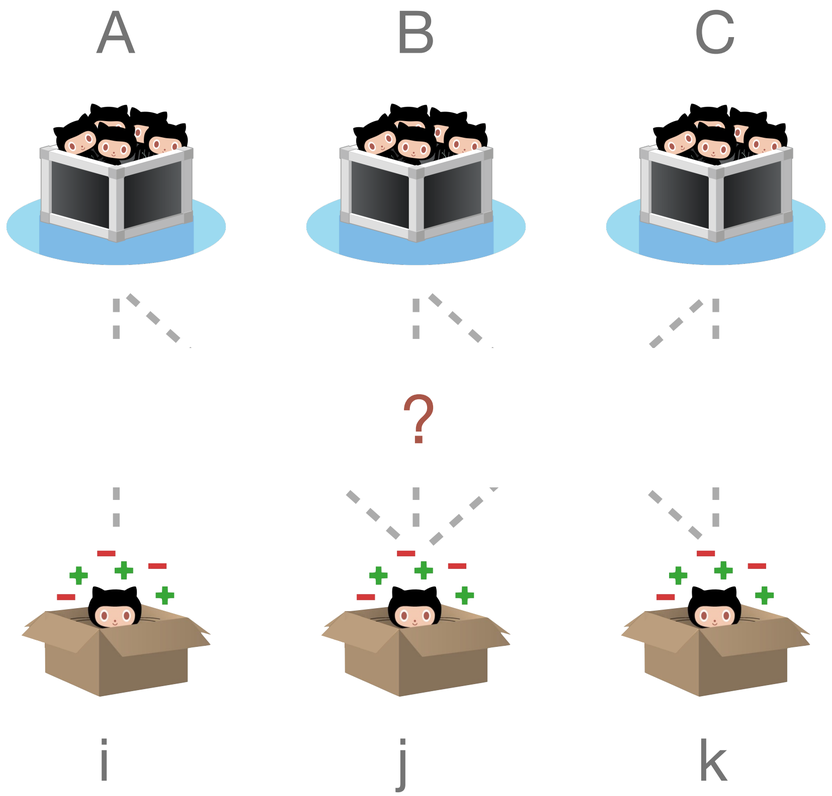


Project A adds a dependency on package j.
New combinations are formed, e.g., (i, j).

How atypical is (i, j)?

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

Counterfactual:

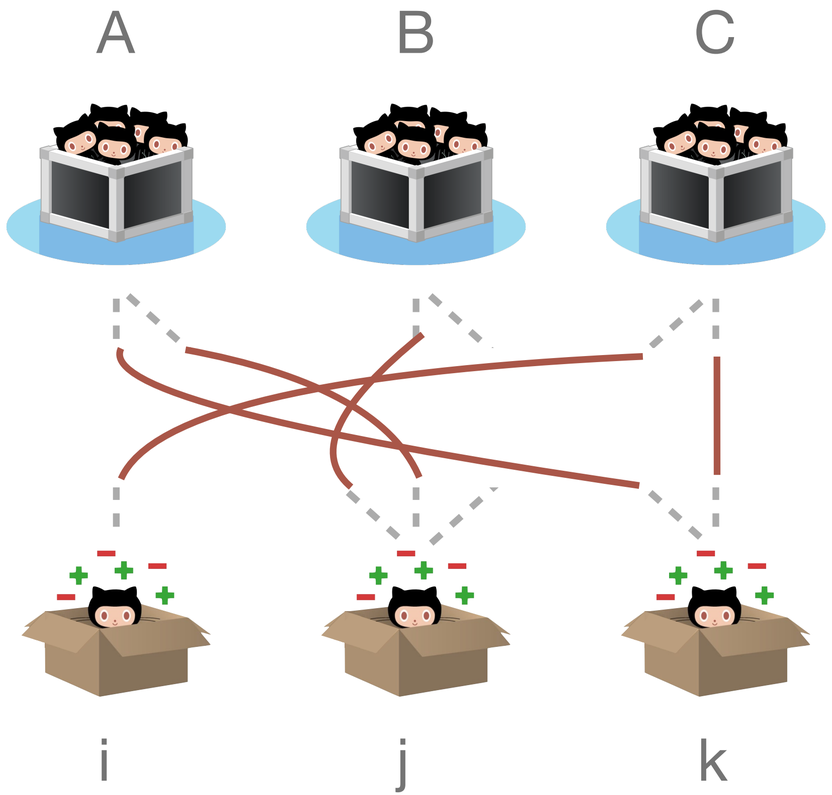


Preserve:

- 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:



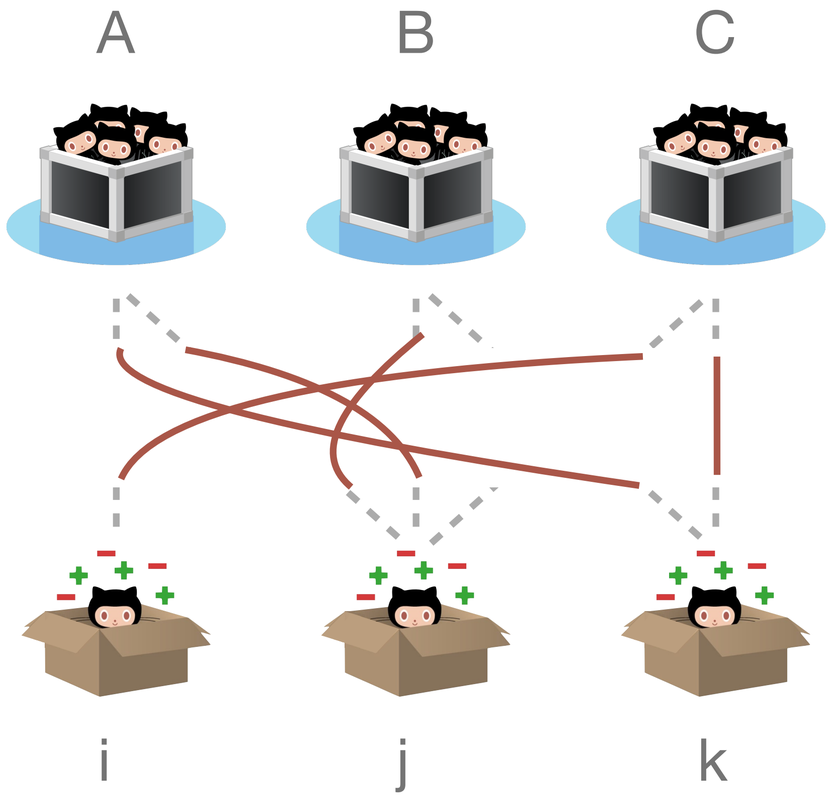
Preserve:

- 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:



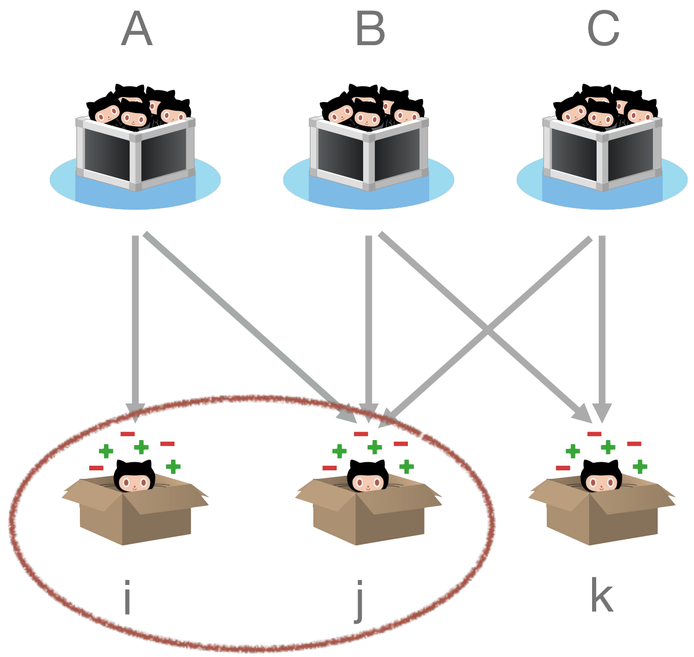
Preserve:

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

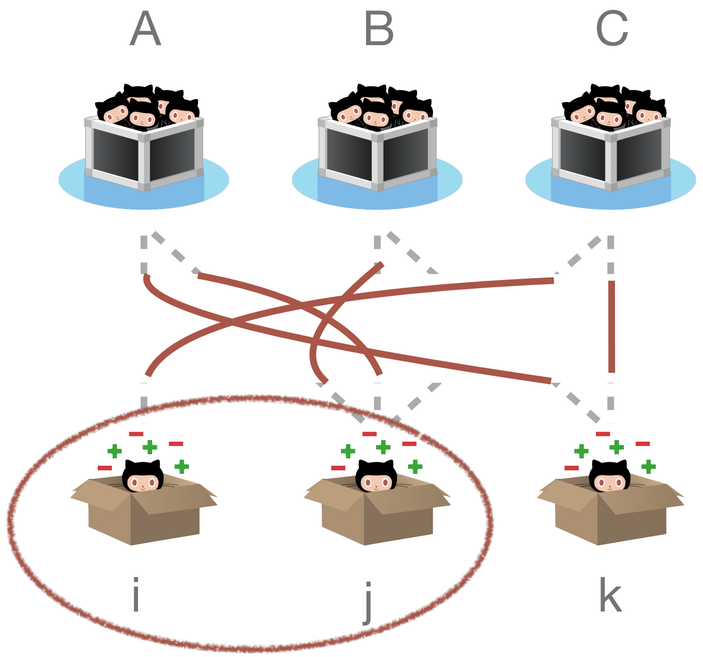
But randomly rewire the network.

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.



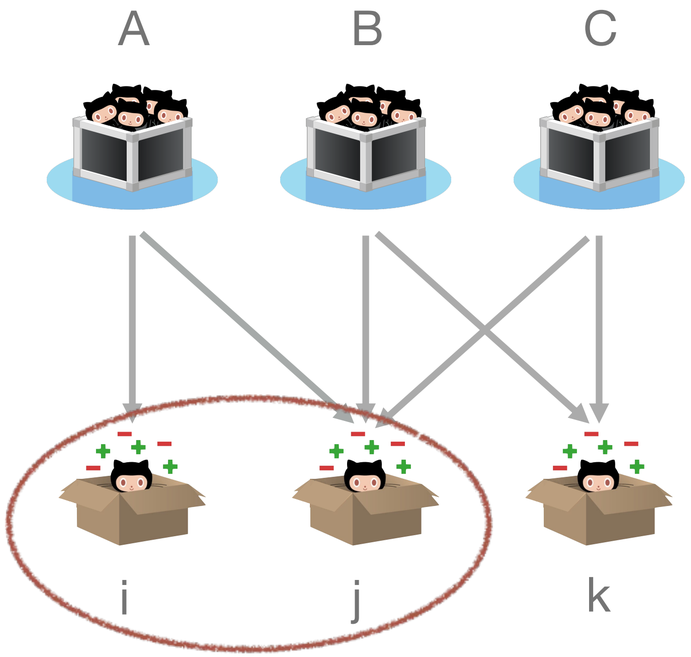
Observed number of times packages i and j appeared together until year t.



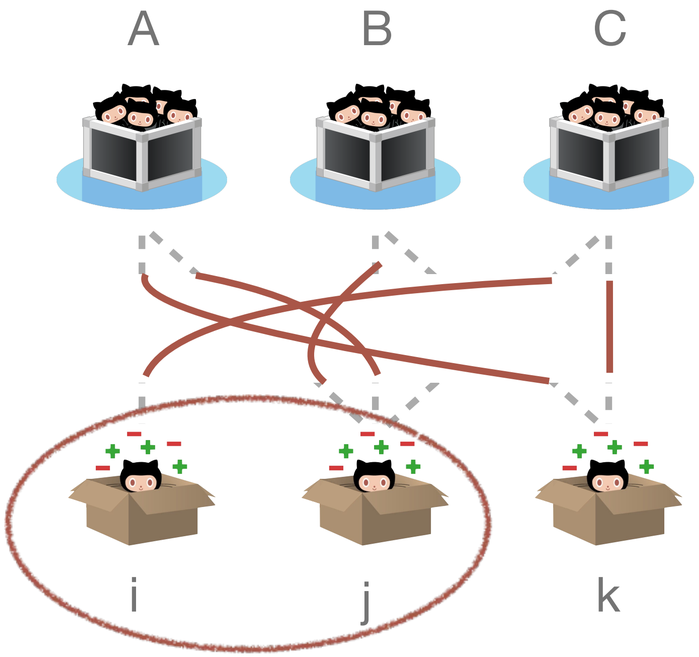
Average (i.e., expected) number of times packages i and j appeared together over N simulations.

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

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

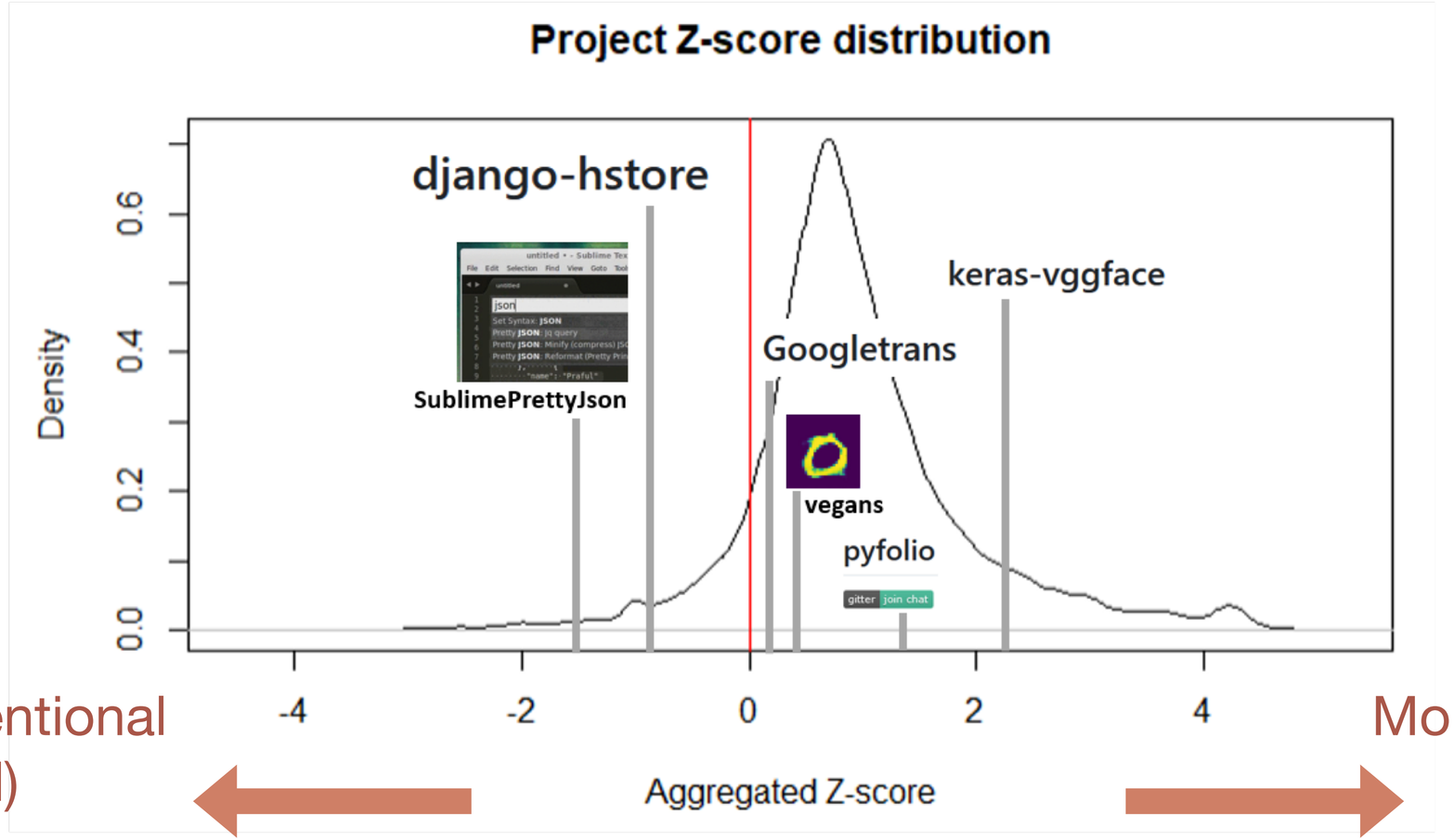
high

⇒ 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.



See paper for more examples and validation

More unconventional (atypical)

More conventional (typical)

Recall our hypotheses

Hyp: Projects that use more atypical combinations of libraries tend to be more popular.

→ Number of GitHub stars by time t

Hyp: More innovative projects tend to be less sustainable.

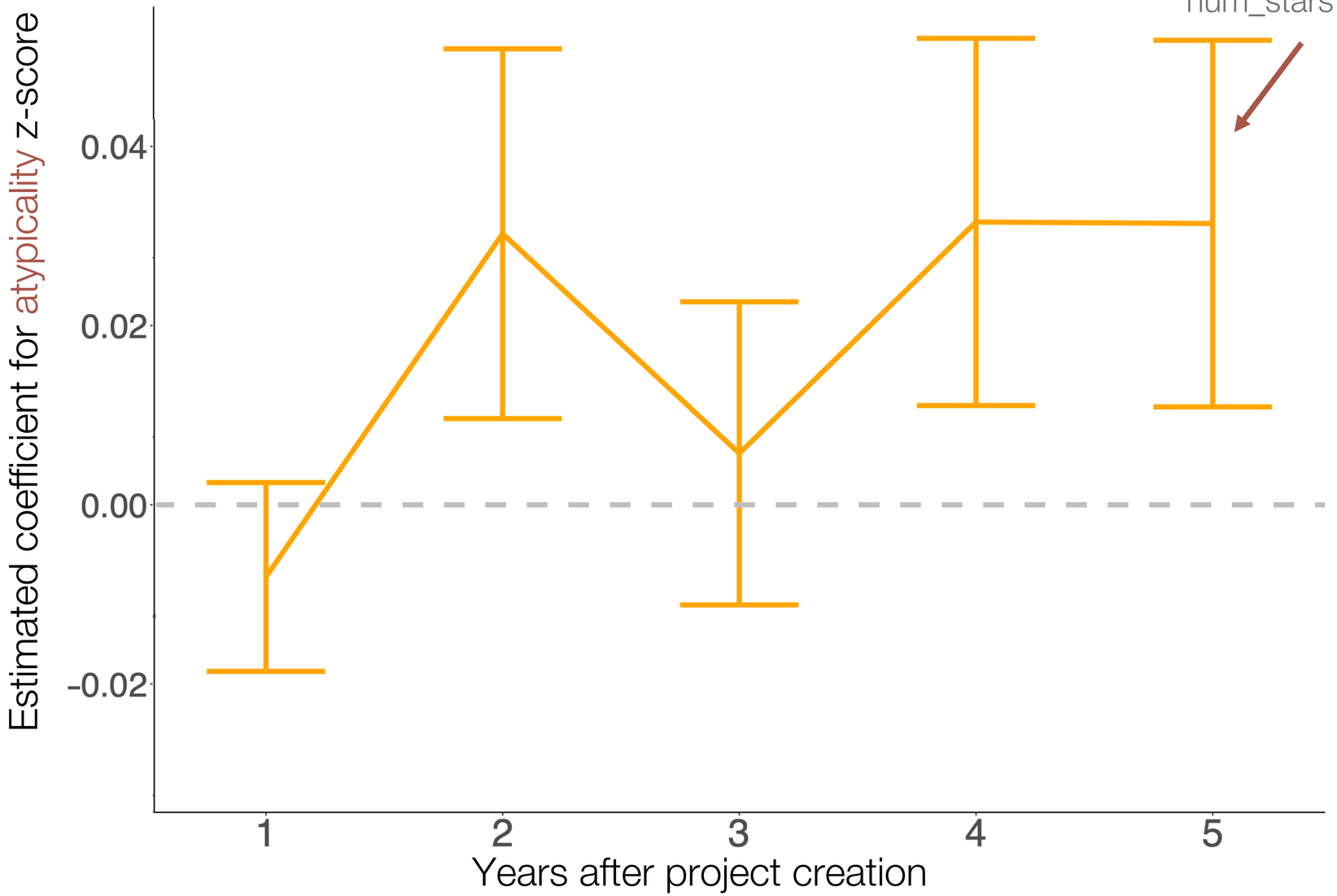
→ Number of new contributors joining by time t

→ Time till project becomes abandonment

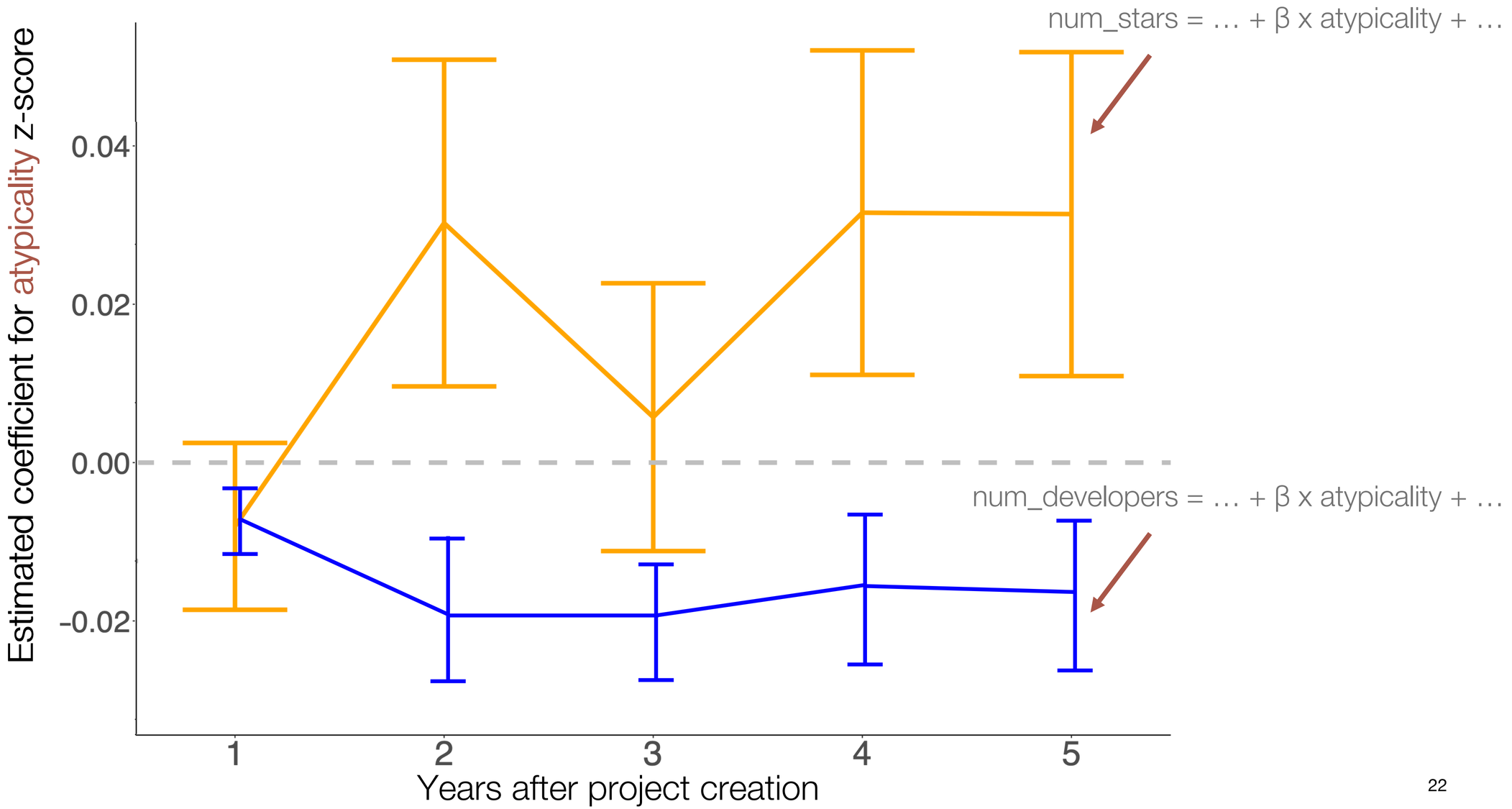
https://www.tasteofhome.com/recipes/chocolate-peanut-butter-pizza/

Atypical (novel) projects tend to have more stars.

num_stars = … + β x atypicality + …

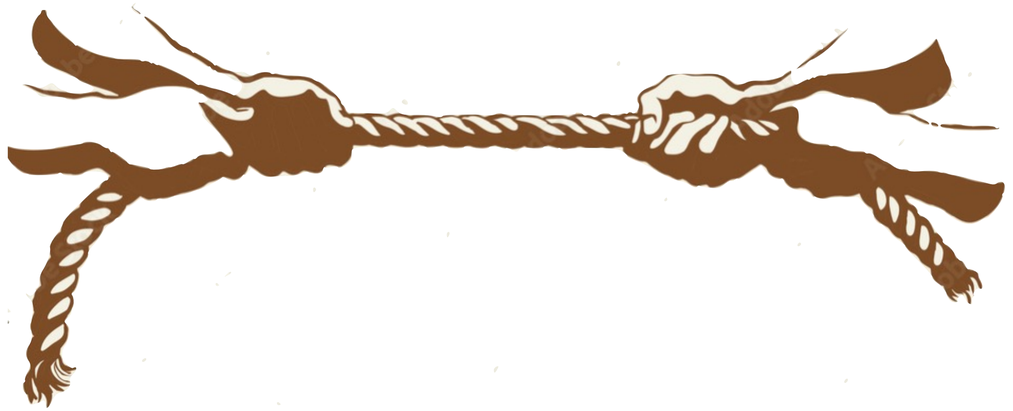


Atypical (novel) projects tend to have smaller teams (and higher probability of becoming abandoned).



Tension between innovation and open source sustainability?

Incentive to create ever-new things

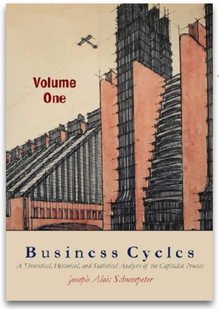


The “grunt work” of maintaining existing systems

- 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?

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[3,4]*

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.

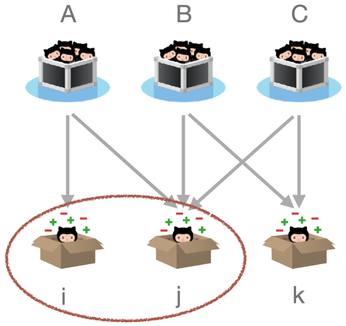
Scientific enterprises are increasingly concerned that research within narrow boundaries is unlikely to be the source of the most fruitful ideas (*1*). Models of creativity emphasize that innovation is spurred through original combinations that spark new insights (*2*–*10*). Current interest in team 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 between exten
binations of k
advantages of
ing is critical t
and impact. H
composition c
can achieve it
In this stuc
search articles
see how prior v
that indicate (i
pers reference
nations of prio
papers based
upon, and (iii)
collaboration.
We conside
ences in the bi
We counted t
pair across all
WOS and corr
to those expec
citation netwo
networks, all
in the WOS w
Carlo algorith
serves the tota
paper and th
counts forwarc
that a paper (
observed netw
randomized ne
the randomizes
we aggregated
respective jour
combinations
over 122 milli
by the 15,613

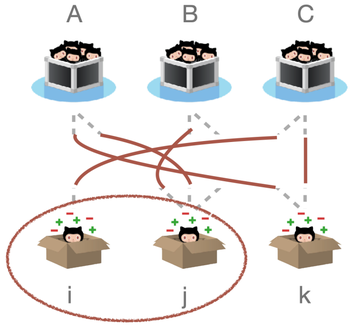
[1]Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, USA. [2]Northwestern Institute on Complex Systems, Northwestern University, 600 Foster, Evanston, IL 60208, USA. [3]Datascope Analytics, 180 West Adams Street, Chicago, IL 60603, USA. [4]National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138, USA.
*Corresponding author. E-mail: bjones@kellogg.northwestern.edu

(Uzzi et al, 2013)

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.

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

Thanks!

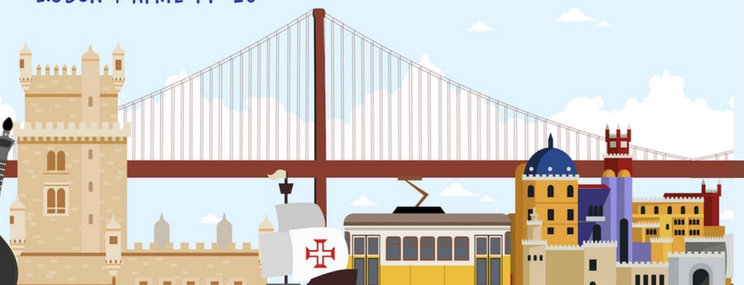
Hongbo Fang
@fang_hongbo

Jim Herbsleb
@jherbsleb

Bogdan Vasilescu
@b_vasilescu

PORTUGAL
LISBON | APRIL 14–20

ICSE 24



STRUDEL

S3D Software and Societal Systems Department

Carnegie Mellon University