The Upper Bound of Information Diffusion in Code Review

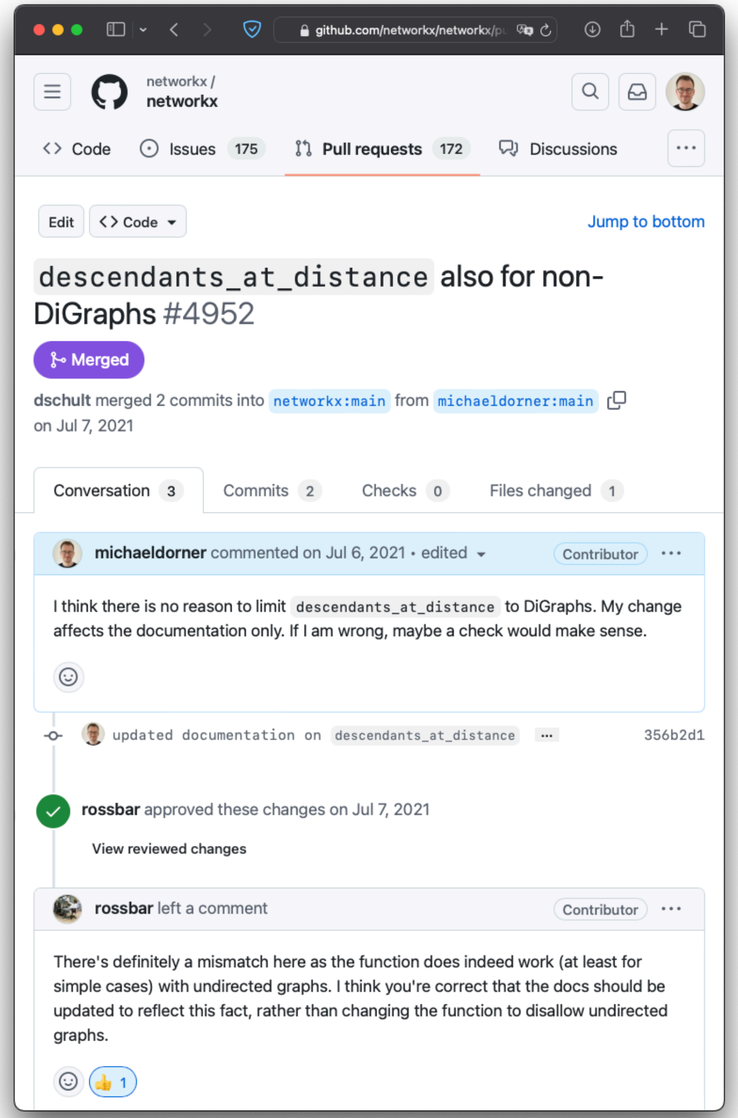
Michael Dorner
Daniel Mendez
Krzysztof Wnuk
Ehsan Zabardast
Jacek Czerwonka


BLACK LIVES MATTER

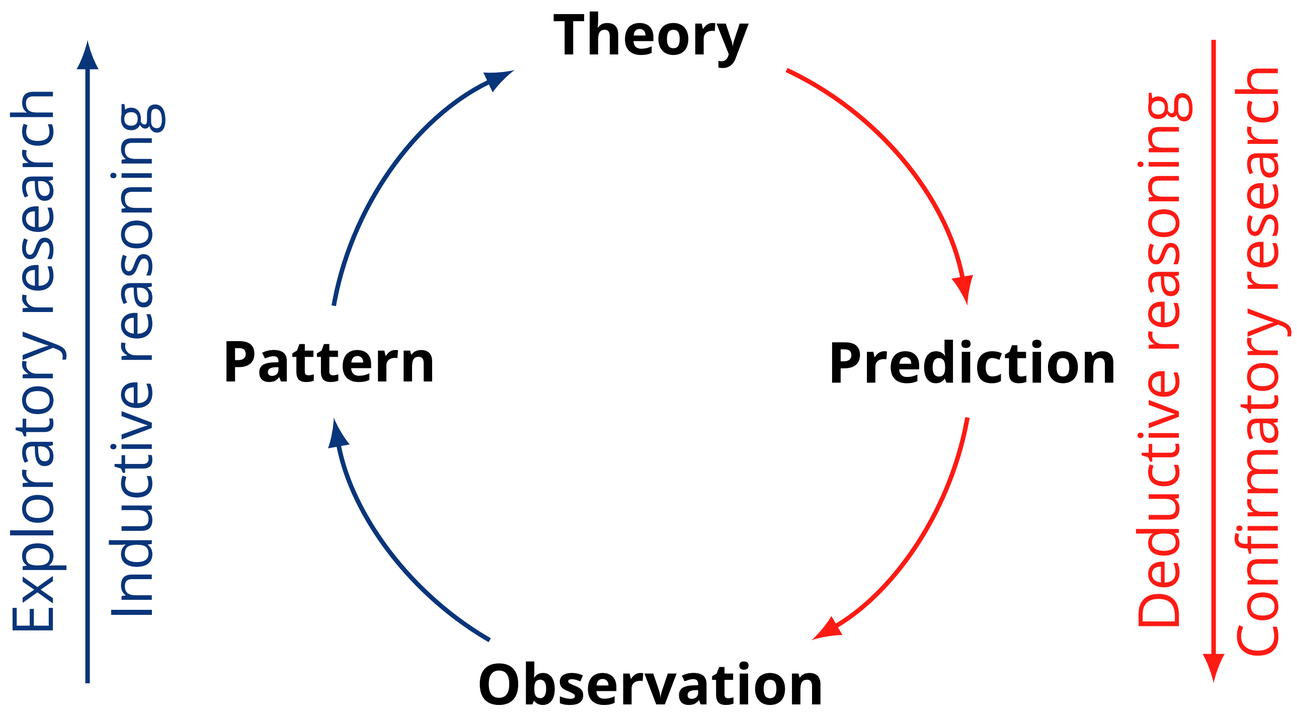
Code Review

- is the discussion around a code change among developers
- is expected to exchange information [1 - 5]

→ **as communication network**



What is missing?



LaTeX source: https://www.michaeldorner.de/posts/exploratory-vs-confirmatory-research/

Our Research Objective

- Provide a confirmatory counterpart to the exploratory prior work
- Evaluate the capability of code review as communication network

Why has it not been done before?

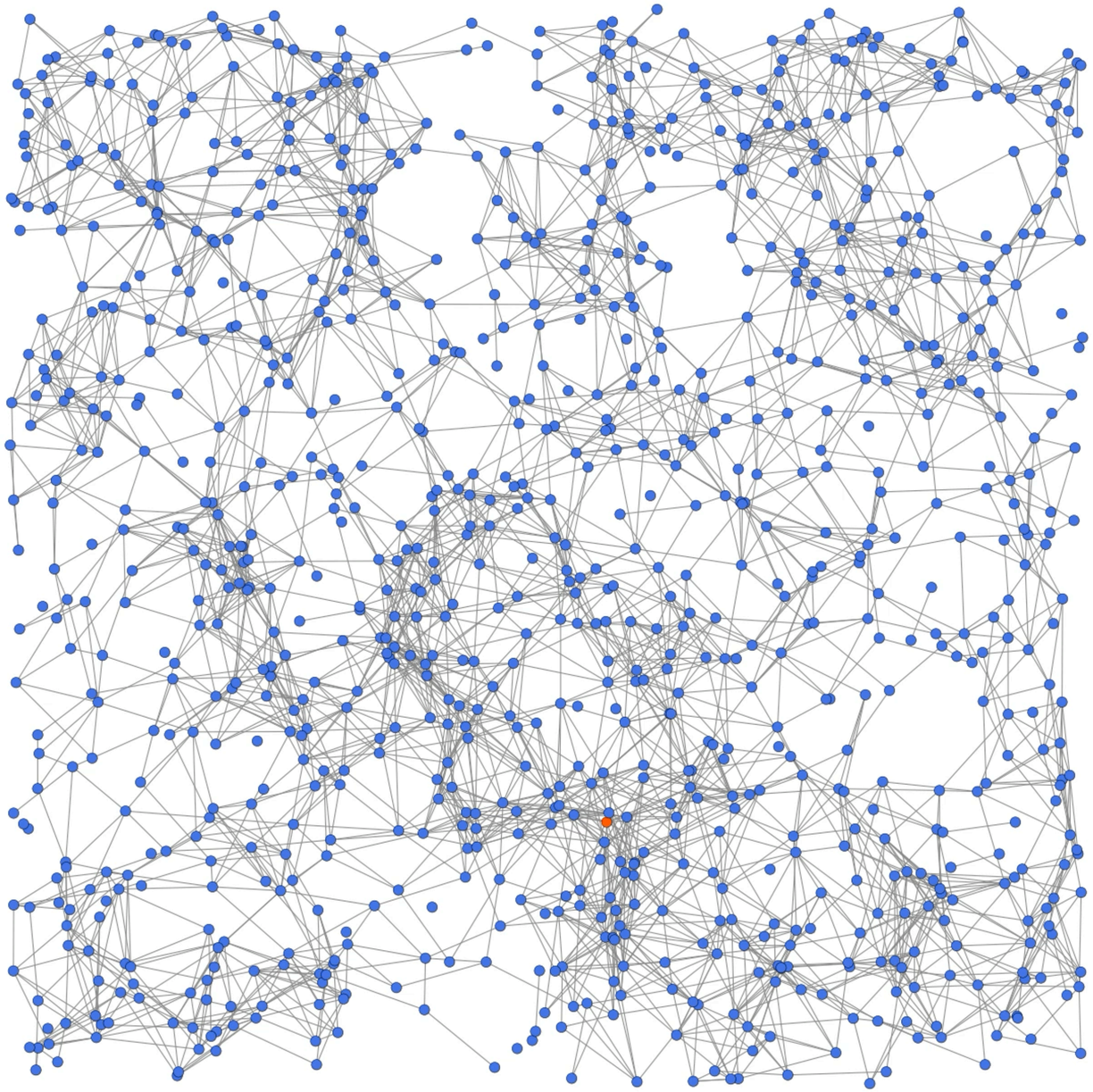
Challenges

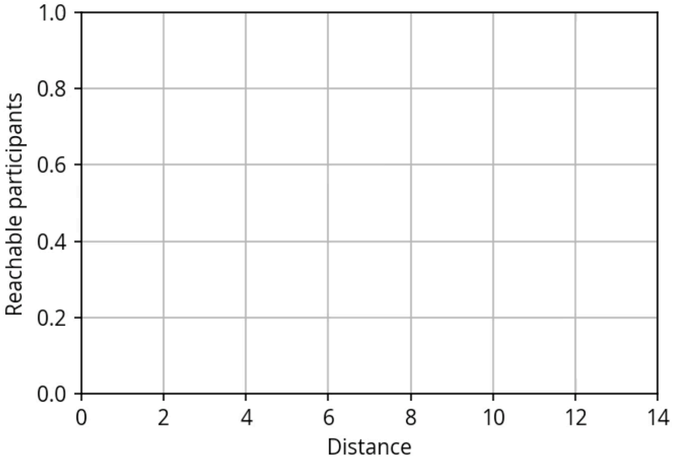
- Confirmatory research usually requires an **experiment** or measurement to falsify the prediction.
- Traditional experiments with whole code review systems (meaning companies) are practically impossible.

How do we tackle the problem?

Method

- We simulate the spread of information and quantify how fast and how far information can spread through code review under best-case assumptions
- An *in-silico* experiment is an experiment performed in a computer simulation (“on silicon”)
  - Useful when traditional experiments are too complex, too expensive, too lengthy, not possible or accessible
  - Make assumptions explicit
  - Fully reproducible
- No actual information spread → simulation




1.0
0.8
0.6
0.4
0.2
0.0
Reachable participants
0
2
4
6
8
10
12
14
Distance

What makes it hard?

Communication in Code Review

- **is not necessarily bilateral:** Multiple participants can exchange information during a code review.
- **is time-dependent:** The communication channels (code reviews) are time-dependent.

Time-Varying	Hypergraphs
can handle time	can handle edges connecting multiple vertices

Minimal distance

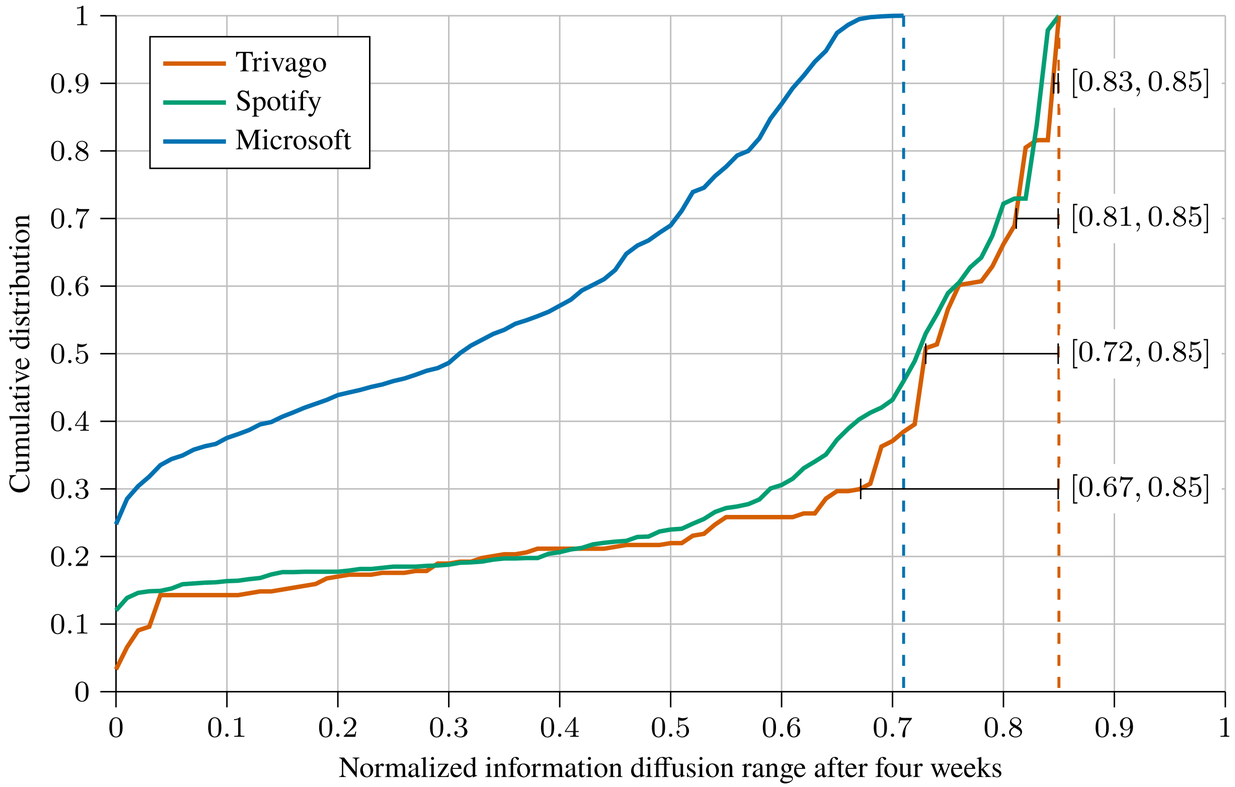
- How **far** information can spread ↔ horizon (connected component)
- How **fast** information can spread ↔ minimal path
  - Topological (shortest)
  - Temporal (fastest)
- Adapted Dijkstra's algorithm for finding all types of minimal distance

Code Review Networks

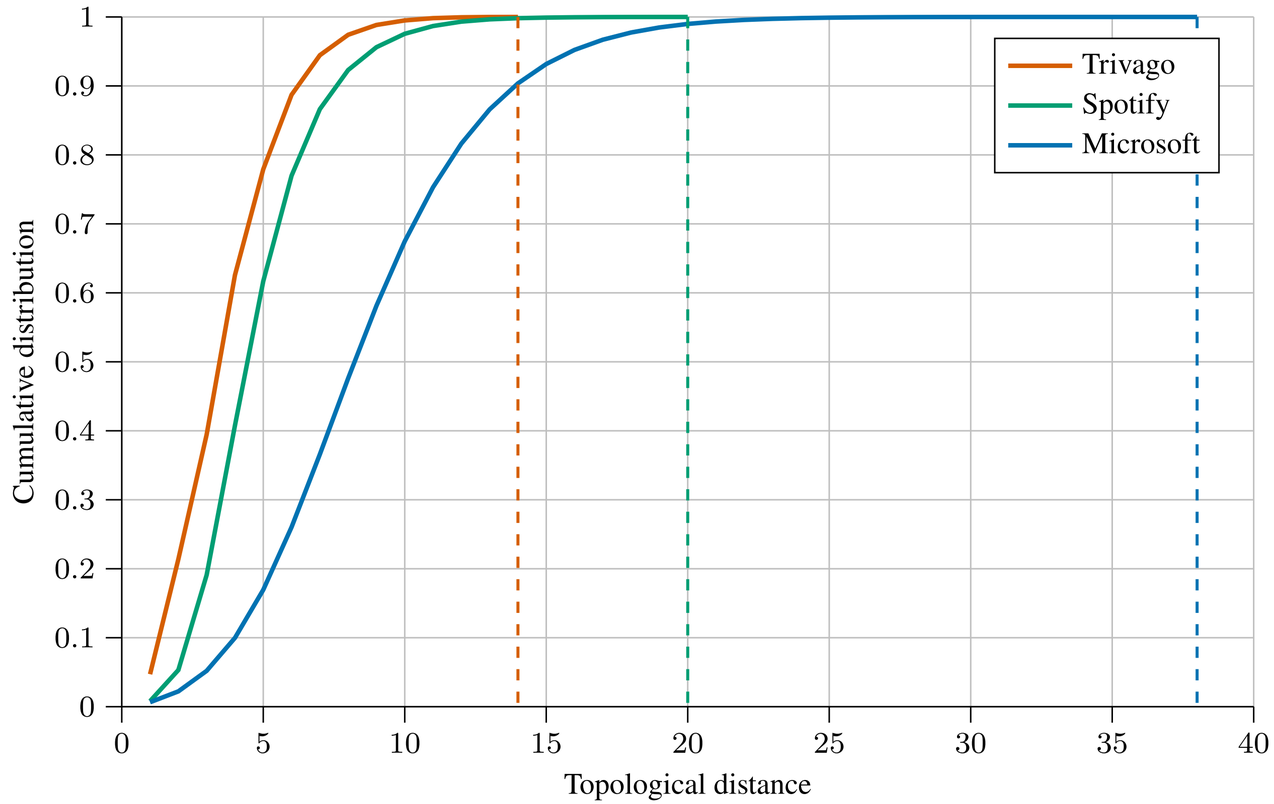
	Code review system size			Tooling
	Classification	**Code reviews**	**Participants**	
Trivago	small	2 442	364	BitBucket
Spotify	mid-sized	22 504	1 730	GitHub
Microsoft	large	309 740	37 103	CodeFlow

Results

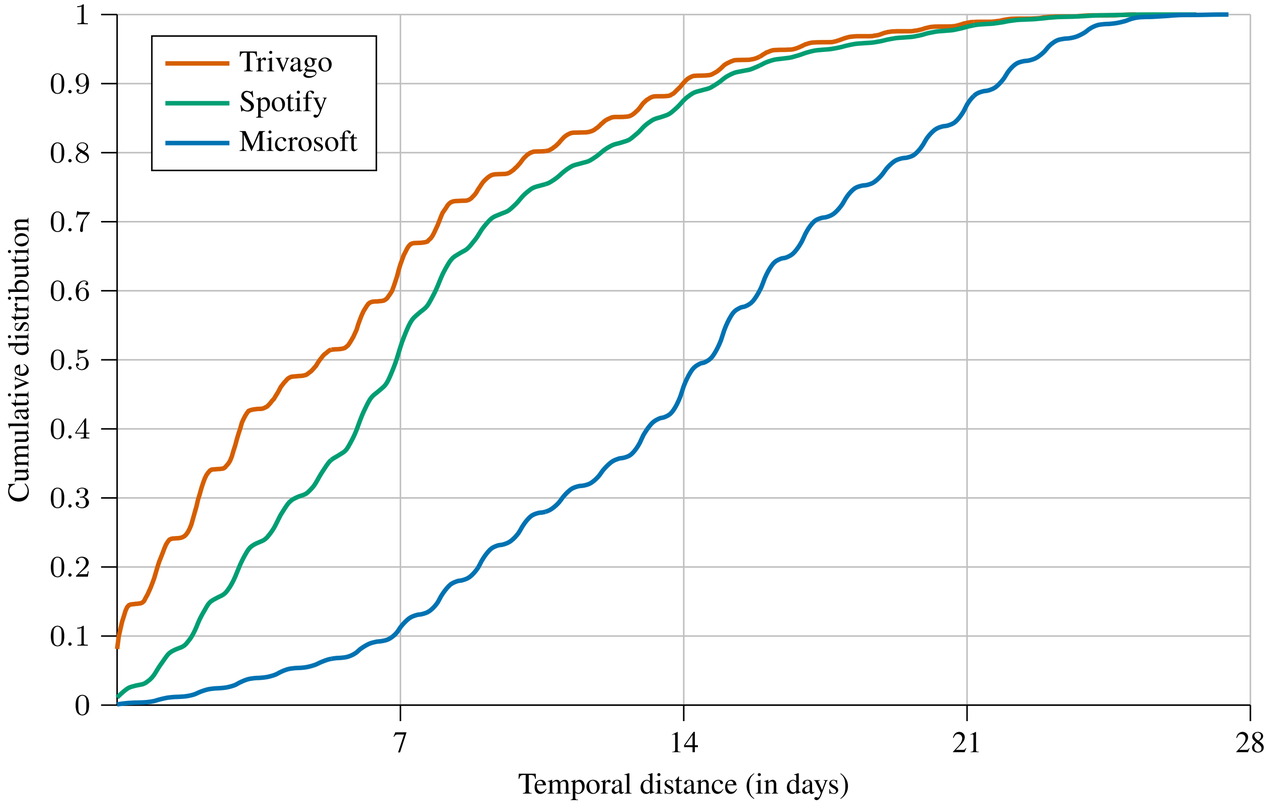
How far can information spread?



How fast can information spread? (1)



How fast can information spread? (2)



Conclusion

Conclusion

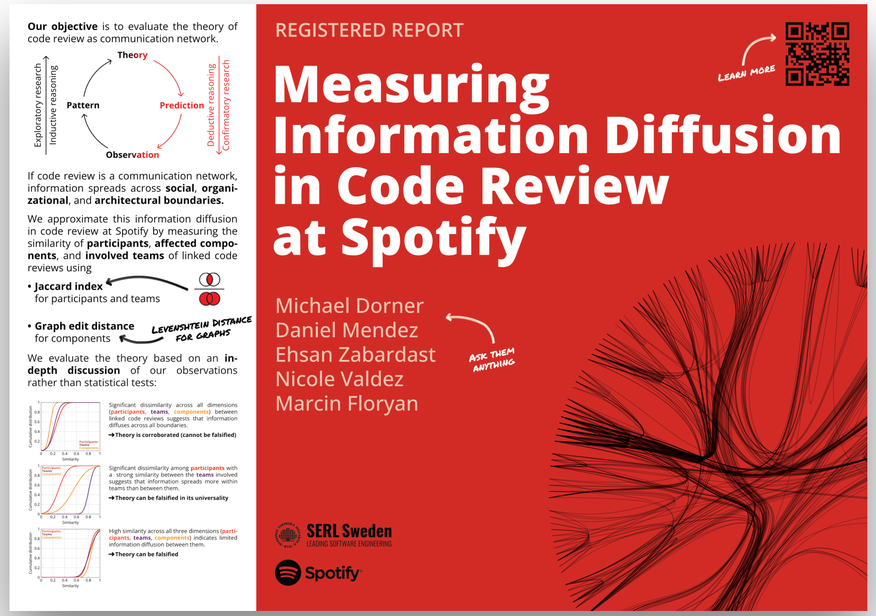
- Code review is capable of spreading information:
  - For **small and mid-sized** code review systems, we observe an **almost identical upper bound** of how far information can spread **relative to the network size**.
  - For **large** code review systems, we observe an **absolute upper bound**.
- A first quantitative confirmation of qualitative findings from prior work

Replication Package

- Empirical Software Engineering (EMSE) open science badge
- All code and data is fully available under MIT on GitHub and Zenodo

What’s next?

- **Measuring** (not simulating) the actual information diffusion in code review
- Compare our findings from industrial code review to open source



Key Take-Aways

- The large the better (regarding information diffusion in code review systems)
- Bots need to be rethought
- Tooling is secondary
- If code review spreads non-trivial information across boundaries, it becomes taxable.

All in one slide

- Information diffusion is the spread of information among humans.
- Code review scales the communication between developers
- Towards understanding code review as a communication network

Got excited?

- Title: “The Upper Bound of Information Diffusion in Code Review”
- Journal: Empirical Software Engineering
- DOI: **10.1007/s10664-024-10442-y**

Empirical Software Engineering (2025) 30:2
https://doi.org/10.1007/s10664-024-10442-y

Check for updates

The upper bound of information diffusion in code review

Michael Dorner[1] · Daniel Mendez[1,2] · Krzysztof Wnuk[1] · Ehsan Zabardast[1] · Jacek Czerwonka[3]

Accepted: 3 January 2024

Abstract

Background Code review, the discussion around a code change among humans, forms a communication network that enables its participants to exchange and spread information. Although reported by qualitative studies, our understanding of the capability of code review as a communication network is still limited.

Objective In this article, we report on a first step towards understanding and evaluating the capability of code review as a communication network by quantifying how fast and how far information can spread through code review: the upper bound of information diffusion in code review.

Method In an *in-silico* experiment, we simulate an artificial information diffusion within large (Microsoft), mid-sized (Spotify), and small code review systems (Trivago) modelled as communication networks. We then measure the minimal topological and temporal distances between the participants to quantify how far and how fast information can spread in code review.

Results An average code review participants in the small and mid-sized code review systems can spread information to between 72 % and 85 % of all code review participants within four weeks independently of network size and tooling; for the large code review systems, we found an absolute boundary of about 11 000 reachable participants. On average (median), information can spread between two participants in code review in less than five hops and less than five days.

Conclusion We found evidence that the communication network emerging from code review scales well and spreads information fast and broadly, corroborating the findings of prior qualitative work. The study lays the foundation for understanding and improving code review as a communication network.

Communicated by: Klaas-Jan Stol

This article belongs to the Topical Collection: *Open Science*

This paper has been awarded the Empirical Software Engineering (EMSE) open science badge.

✉ Michael Dorner
michael.dorner@bth.se

[1] Blekinge Institute of Technology, Karlskrona, Sweden

[2] Fortiss, München, Germany

[3] Microsoft, Seattle, USA

Published online: 17 October 2024

Springer