Rust Analytics for Software Heritage

Challenges and results

Tommaso Fontana**, Sebastiano Vigna,** Stefano Zacchiroli

Partially supported by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU, and by project ANR COREGRAPHIE, grant ANR-20-CE23-0002 of the French Agence Nationale de la Recherche

The Laboratory for Web Algorithmics

The Laboratory for Web Algorithmics

- A laboratory established in the late 90s at the Università degli Studi di Milano to attack specific algorithmic problems related to the web

The Laboratory for Web Algorithmics

- A laboratory established in the late 90s at the Università degli Studi di Milano to attack specific algorithmic problems related to the web
- Expanded later into the study of social networks

The Laboratory for Web Algorithmics

- A laboratory established in the late 90s at the Università degli Studi di Milano to attack specific algorithmic problems related to the web
- Expanded later into the study of social networks
- Besides scientific publications, `several` open-source Java projects

The Laboratory for Web Algorithmics

- A laboratory established in the late 90s at the Università degli Studi di Milano to attack specific algorithmic problems related to the web
- Expanded later into the study of social networks
- Besides scientific publications, `several` open-source Java projects
- For example, `fastutil` is currently used by 775 open-source Java projects in the Maven Central repository

The Laboratory for Web Algorithmics

- A laboratory established in the late 90s at the Università degli Studi di Milano to attack specific algorithmic problems related to the web
- Expanded later into the study of social networks
- Besides scientific publications, `several` open-source Java projects
- For example, `fastutil` is currently used by 775 open-source Java projects in the Maven Central repository
- We aim at strong theoretical results that have an actual impact on the industry

The Laboratory for Web Algorithmics

- A laboratory established in the late 90s at the Università degli Studi di Milano to attack specific algorithmic problems related to the web
- Expanded later into the study of social networks
- Besides scientific publications, `several` open-source Java projects
- For example, `fastutil` is currently used by 775 open-source Java projects in the Maven Central repository
- We aim at strong theoretical results that have an actual impact on the industry
- We also harvest data such as web snapshots and make them publicly available for researchers

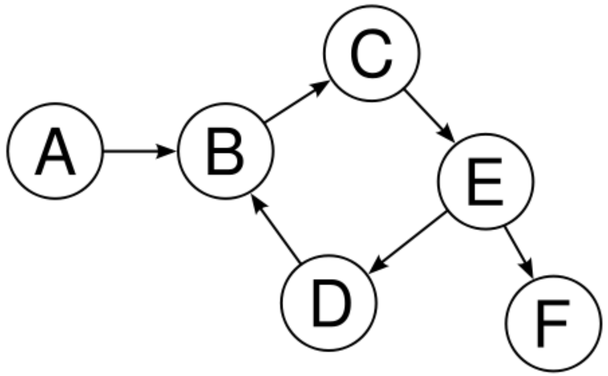
Graph Analytics at Scale

Graph Analytics at Scale

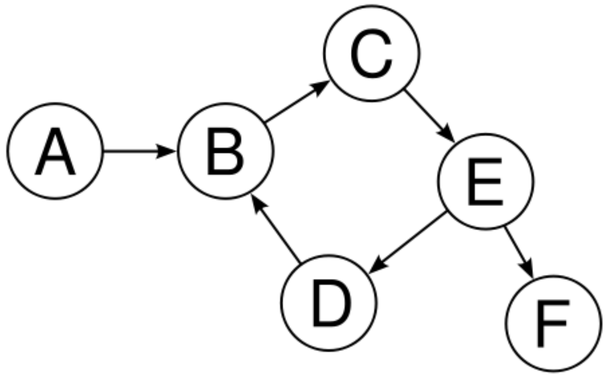
- Today we found very large graphs all over the place

Graph Analytics at Scale

- Today we found very large graphs all over the place



Graph Analytics at Scale



- Today we found very large graphs all over the place
- Web snapshots, social networks, biological graphs, …

Graph Analytics at Scale

- Today we found very large graphs all over the place
- Web snapshots, social networks, biological graphs, …

Graph Analytics at Scale

- Today we found very large graphs all over the place
- Web snapshots, social networks, biological graphs, …

Graph Analytics at Scale

- Today we found very large graphs all over the place
- Web snapshots, social networks, biological graphs, …
- Very large graphs require new approaches

Graph Analytics at Scale

- Today we found very large graphs all over the place
- Web snapshots, social networks, biological graphs, …
- Very large graphs require new approaches
- Standard representations in main memory are either impossible (graph too large) or impossibly expensive (many TB of core memory)

Graph Analytics at Scale

- Today we found very large graphs all over the place
- Web snapshots, social networks, biological graphs, …
- Very large graphs require new approaches
- Standard representations in main memory are either impossible (graph too large) or impossibly expensive (many TB of core memory)
- Distributed approaches spend an impossible amount of time distributing data among nodes

Graph Analytics at Scale

- Today we found very large graphs all over the place
- Web snapshots, social networks, biological graphs, …
- Very large graphs require new approaches
- Standard representations in main memory are either impossible (graph too large) or impossibly expensive (many TB of core memory)
- Distributed approaches spend an impossible amount of time distributing data among nodes
- What can we do?

The WebGraph Framework

The WebGraph Framework

- An open-source framework for data analytics on very large graphs

The WebGraph Framework

- An open-source framework for data analytics on very large graphs
- One of the most long-lived projects of this kind (>20 years!)

The WebGraph Framework

- An open-source framework for data analytics on very large graphs
- One of the most long-lived projects of this kind (>20 years!)
- Hundreds of publications in major conferences and journals using it (>1500 references)

The WebGraph Framework

- An open-source framework for data analytics on very large graphs
- One of the most long-lived projects of this kind (>20 years!)
- Hundreds of publications in major conferences and journals using it (>1500 references)
- In 2011 news went around the world: Facebook had four degrees of separation

The WebGraph Framework

- An open-source framework for data analytics on very large graphs
- One of the most long-lived projects of this kind (>20 years!)
- Hundreds of publications in major conferences and journals using it (>1500 references)
- In 2011 news went around the world: Facebook had four degrees of separation


HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS

The New York Times

Business Day
Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH

Separating You and Me? 4.74 Degrees

By JOHN MARKOFF and SOMINI SENGUPTA
Published: November 21, 2011

The world is even smaller than you thought.

RECOMMEND
TWITTER

The WebGraph Framework

- An open-source framework for data analytics on very large graphs
- One of the most long-lived projects of this kind (>20 years!)
- Hundreds of publications in major conferences and journals using it (>1500 references)
- In 2011 news went around the world: Facebook had four degrees of separation

The WebGraph Framework

- An open-source framework for data analytics on very large graphs
- One of the most long-lived projects of this kind (>20 years!)
- Hundreds of publications in major conferences and journals using it (>1500 references)
- In 2011 news went around the world: Facebook had four degrees of separation
- The measurement was performed at Facebook in collaboration with our group using WebGraph (at that time, 700M nodes,

The WebGraph Framework

- An open-source framework for data analytics on very large graphs
- One of the most long-lived projects of this kind (>20 years!)
- Hundreds of publications in major conferences and journals using it (>1500 references)
- In 2011 news went around the world: Facebook had four degrees of separation
- The measurement was performed at Facebook in collaboration with our group using WebGraph (at that time, 700M nodes,
- Note: the data structures currently used by Facebook for storing the friendship graph were suggested by us following this interaction

Software Heritage History Graph

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Software Heritage History Graph

- One of the largest graphs of human activity available

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Software Heritage History Graph

- One of the largest graphs of human activity available
- 34 billion nodes, 517 billion arcs (September 2023)

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Software Heritage History Graph

- One of the largest graphs of human activity available
- 34 billion nodes, 517 billion arcs (September 2023)
- Constantly increasing

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Software Heritage History Graph

- One of the largest graphs of human activity available
- 34 billion nodes, 517 billion arcs (September 2023)
- Constantly increasing
- Represented by WebGraph in 176GB instead of 4TB!

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Software Heritage History Graph

- One of the largest graphs of human activity available
- 34 billion nodes, 517 billion arcs (September 2023)
- Constantly increasing
- Represented by WebGraph in 176GB instead of 4TB!
- The current WebGraph-based pipeline for graph analytics was born out of a collaboration between Inria and the Università degli Studi di Milano

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Software Heritage History Graph

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

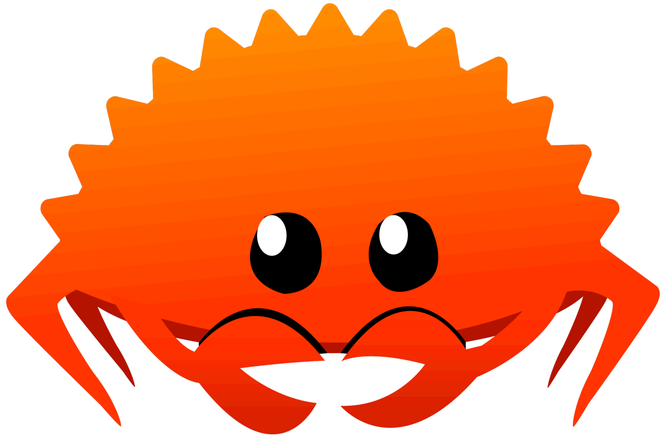
- One of the largest graphs of human activity available
- 34 billion nodes, 517 billion arcs (September 2023)
- Constantly increasing
- Represented by WebGraph in 176GB instead of 4TB!
- The current WebGraph-based pipeline for graph analytics was born out of a collaboration between Inria and the Università degli Studi di Milano
- Storing explicitly the graph makes it possible to perform provenance analysis, plagiarism detection, clone detection, etc., at an unprecedented scale

Software Heritage History Graph

- One of the largest graphs of human activity available
- 34 billion nodes, 517 billion arcs (September 2023)
- Constantly increasing
- Represented by WebGraph in 176GB instead of 4TB!
- The current WebGraph-based pipeline for graph analytics was born out of a collaboration between Inria and the Università degli Studi di Milano
- Storing explicitly the graph makes it possible to perform provenance analysis, plagiarism detection, clone detection, etc., at an unprecedented scale
- Still, Java started to get in the way

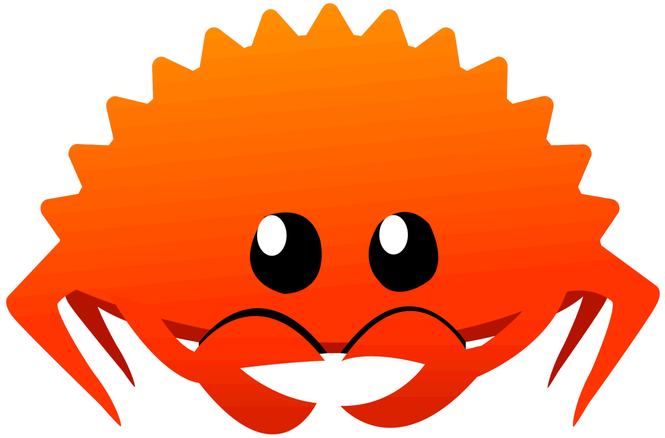
Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Why Rust



Why Rust

- A new, high-performance, safe language



Why Rust

- A new, high-performance, safe language
- Memory safe (as Java), but with zero-cost abstractions

Why Rust

- A new, high-performance, safe language
- Memory safe (as Java), but with zero-cost abstractions
- Arrays as large as memory allows

Why Rust

- A new, high-performance, safe language
- Memory safe (as Java), but with zero-cost abstractions
- Arrays as large as memory allows
- Fine-grained access to OS facilities

Why Rust

- A new, high-performance, safe language
- Memory safe (as Java), but with zero-cost abstractions
- Arrays as large as memory allows
- Fine-grained access to OS facilities
- Modern, functional-inspired idiomatic programming

Why Rust

- A new, high-performance, safe language
- Memory safe (as Java), but with zero-cost abstractions
- Arrays as large as memory allows
- Fine-grained access to OS facilities
- Modern, functional-inspired idiomatic programming
- First language after C to be integrated in the Linux kernel

Why Rust

- A new, high-performance, safe language
- Memory safe (as Java), but with zero-cost abstractions
- Arrays as large as memory allows
- Fine-grained access to OS facilities
- Modern, functional-inspired idiomatic programming
- First language after C to be integrated in the Linux kernel
- Rapidly adopted by the industry (Google, Facebook, …)

Results

Results

- Collaboration between Inria / Télécom Paris and the Università degli Studi di Milano

Results

- Collaboration between Inria / Télécom Paris and the Università degli Studi di Milano
- > 100,000 committed lines of code in about eight months, > 50,000 present

Results

- Collaboration between Inria / Télécom Paris and the Università degli Studi di Milano
- > 100,000 committed lines of code in about eight months, > 50,000 present
- New graph-analytics pipeline (still some steps to replace)

Results

- Collaboration between Inria / Télécom Paris and the Università degli Studi di Milano
- > 100,000 committed lines of code in about eight months, > 50,000 present
- New graph-analytics pipeline (still some steps to replace)
- Open to new compression models (hardwired in the Java version)

Results

- Collaboration between Inria / Télécom Paris and the Università degli Studi di Milano
- > 100,000 committed lines of code in about eight months, > 50,000 present
- New graph-analytics pipeline (still some steps to replace)
- Open to new compression models (hardwired in the Java version)
- More predictable performance, and almost three times faster!

Results

- Collaboration between Inria / Télécom Paris and the Università degli Studi di Milano
- > 100,000 committed lines of code in about eight months, > 50,000 present
- New graph-analytics pipeline (still some steps to replace)
- Open to new compression models (hardwired in the Java version)
- More predictable performance, and almost three times faster!
- Many satellite open-source projects released to the community

Results

- Collaboration between Inria / Télécom Paris and the Università degli Studi di Milano
- > 100,000 committed lines of code in about eight months, > 50,000 present
- New graph-analytics pipeline (still some steps to replace)
- Open to new compression models (hardwired in the Java version)
- More predictable performance, and almost three times faster!
- Many satellite open-source projects released to the community
- Ready for the future growth of Software Heritage