R012:Techniques for complex analysis of contemporary data

Jakub Peschel, Michal Batko, Pavel Zezula

Masaryk University

July 30, 2020

Overview

Motivation

- Data growth and complexity
- Highly specialised tools vs. General reporting tools
- Complex analytical processes

Source	Users (aprox.)	Data Exchange (aprox.)
Facebook	2.6 billion users	4 mil. / sec.
Gmail	1.5 billion accounts	2. mil. / sec.
Instagram	1 billion users	1.1 thous. / sec

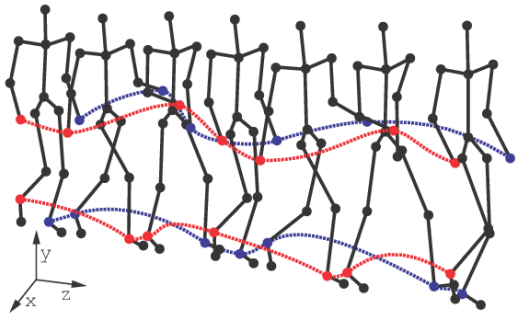
Table: Approximate sizes of different sources of data

Motivation

- Data growth and complexity
- Highly specialised tools vs. General reporting tools
- Complex analytical processes

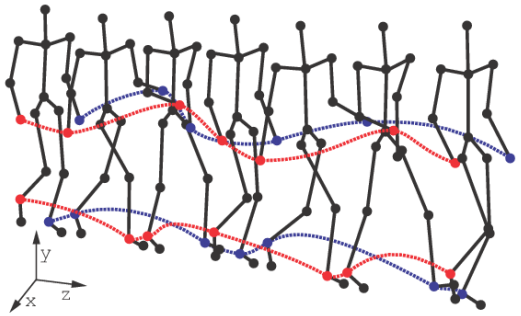
Motivation

- Data growth and complexity
- Highly specialised tools vs. General reporting tools
- Complex analytical processes



Motivation

- Data growth and complexity
- Highly specialised tools vs. General reporting tools
- Complex analytical processes



Motivation

- Data growth and complexity
- Highly specialised tools vs. General reporting tools
- Complex analytical processes

orange

jupyter

spmf

Motivation

- Data growth and complexity
- Highly specialised tools vs. General reporting tools
- Complex analytical processes

orange

jupyter

spmf

Goal: Task-oriented analysis

- Complex analytical operations
- Focus on task vs. focus on implementation
- Proof-of-concept system

Goal: Task-oriented analysis

- Complex analytical operations
- Focus on task vs. focus on implementation
- Proof-of-concept system
- Community mining
  - Core + surroundings

Goal: Task-oriented analysis

- Complex analytical operations
- Focus on task vs. focus on implementation
- Proof-of-concept system

- Community mining
  - Core + surroundings

Goal: Task-oriented analysis

- Complex analytical operations
- Focus on task vs. focus on implementation
- Proof-of-concept system

- Community mining
  - Core + surroundings

Goal: Task-oriented analysis

- Complex analytical operations
- Focus on task vs. focus on implementation
- Proof-of-concept system

- Community mining
  - Core + surroundings
  - Frequent item-set mining + Similarity search

Goal: Task-oriented analysis

- Complex analytical operations
- Focus on task vs. focus on implementation
- Proof-of-concept system
  - Universal storage
- Community mining
  - Core + surroundings
  - Frequent item-set mining + Similarity search

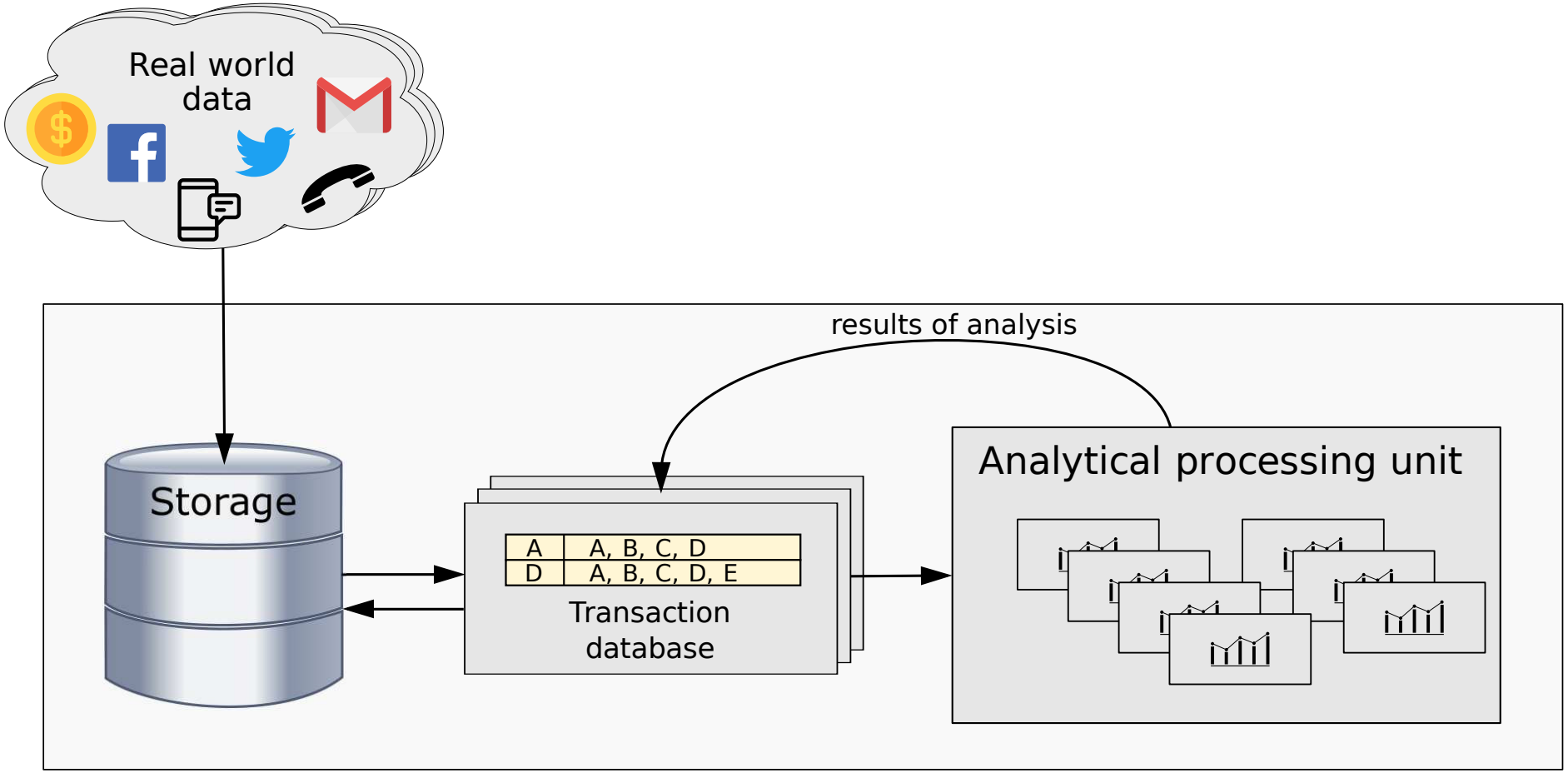
Goal: Task-oriented analysis

- Complex analytical operations
- Focus on task vs. focus on implementation
- Proof-of-concept system
  - Universal storage
  - Basic analytical operations
- Community mining
  - Core + surroundings
  - Frequent item-set mining + Similarity search

Goal: Task-oriented analysis

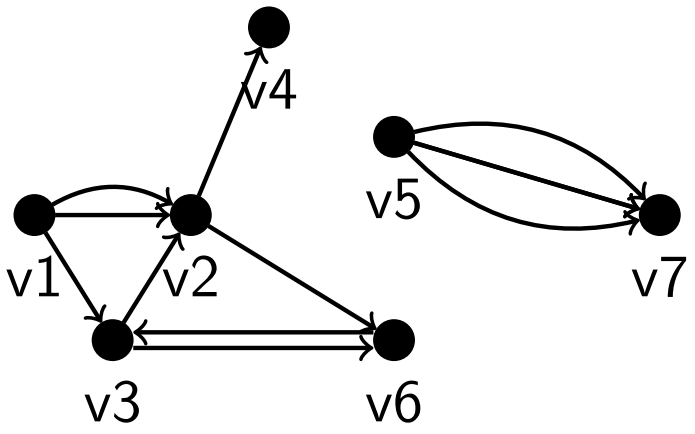
- Complex analytical operations
- Focus on task vs. focus on implementation
- Proof-of-concept system
  - Universal storage
  - Basic analytical operations
  - Chaining of the operations
- Community mining
  - Core + surroundings
  - Frequent item-set mining + Similarity search

ADAMiSS: Overview



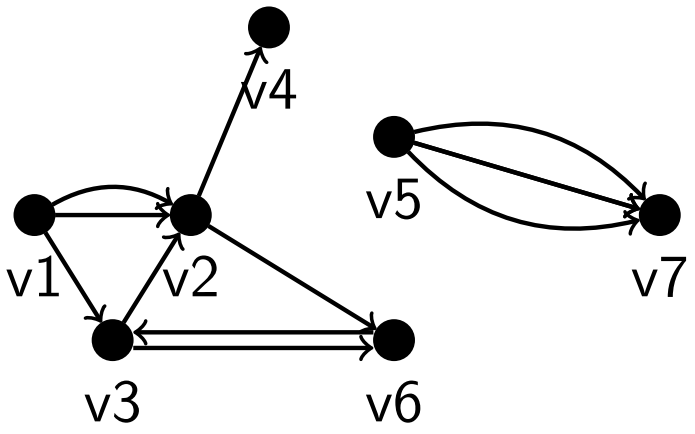
Storage - Graph representation

- Multigraph representation
  - Capable capturing most of the data



Storage - Graph representation

- Multigraph representation
  - Capable capturing most of the data
  - Attributes store additional information

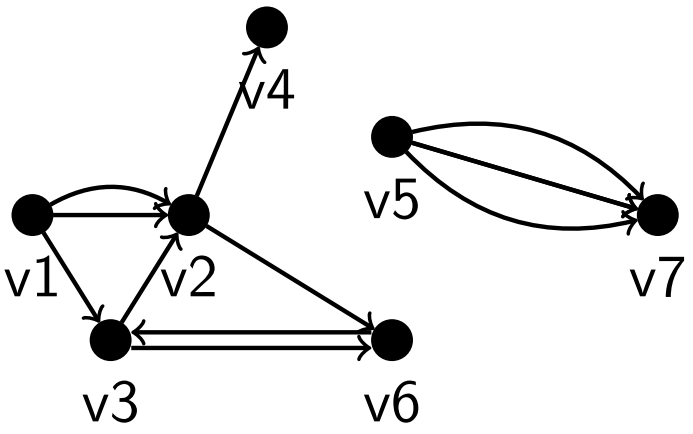


Example

- Nodes: accounts, images, videos
- Edges: e-mails, occurrences

Storage - Graph representation

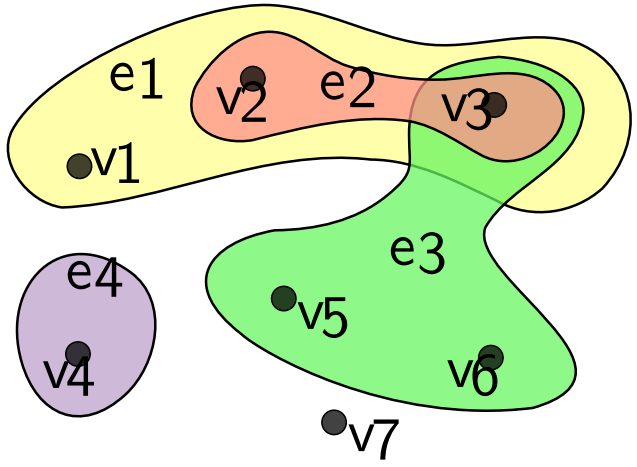
- Multigraph representation
  - Capable capturing most of the data
  - Attributes store additional information



Example

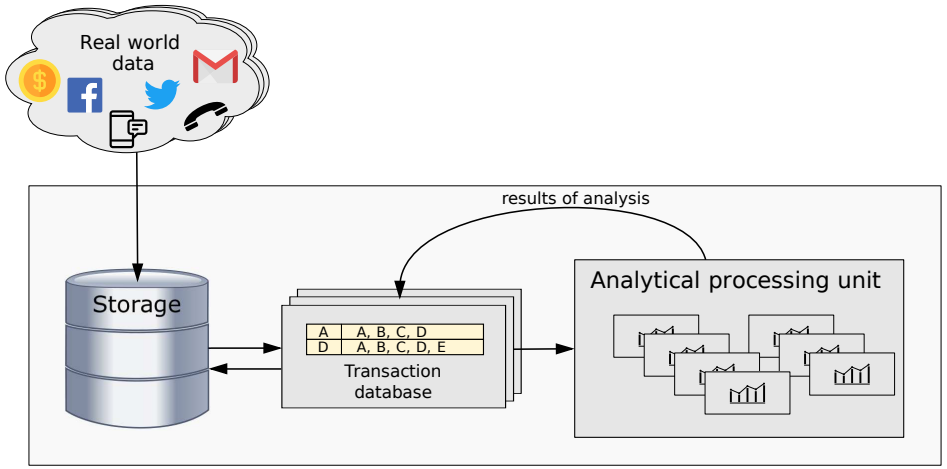
- Nodes: accounts, images, videos
- Edges: e-mails, occurrences

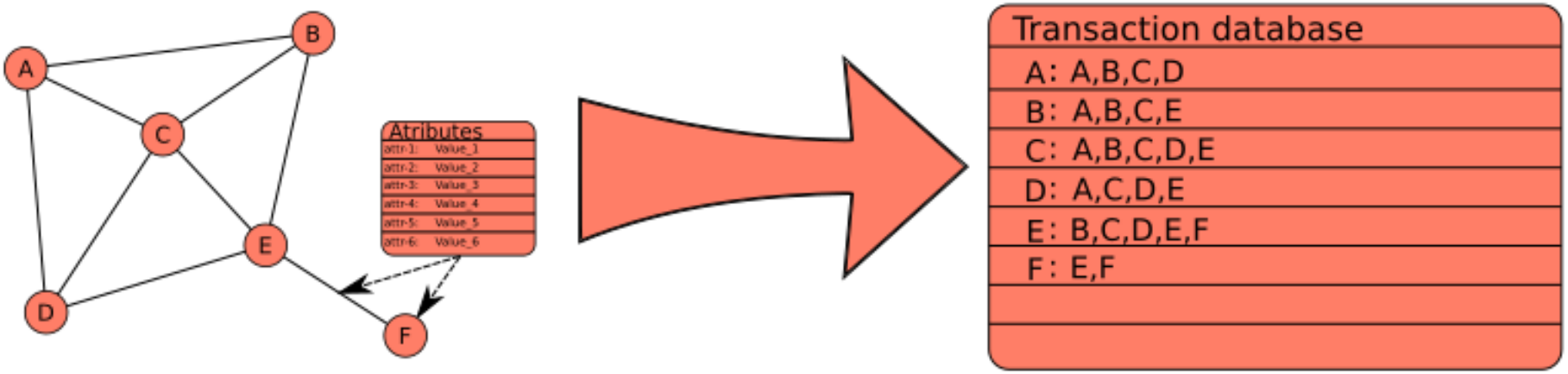
- Is hypergraph a better representation?
  - Standard operations are too complex
  - Attributes can partially store such information



Transaction DB

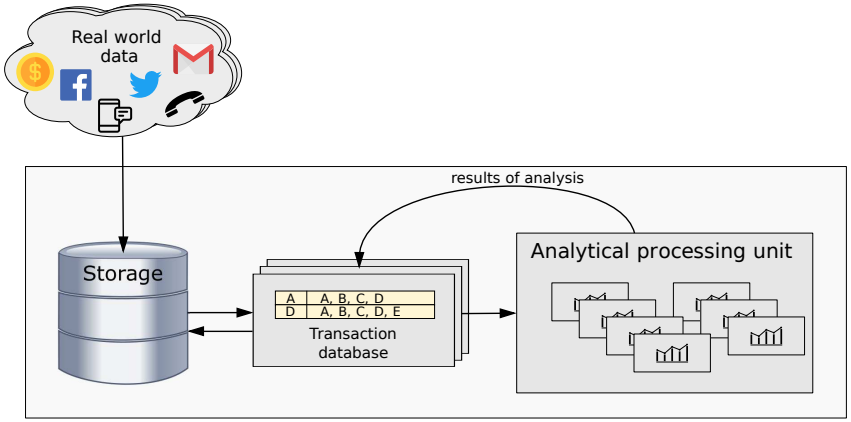
- Flat structure
- Unified for analytical operators
- Transformed from graph
- Filtration based on
  - Properties of graph
  - Attributes





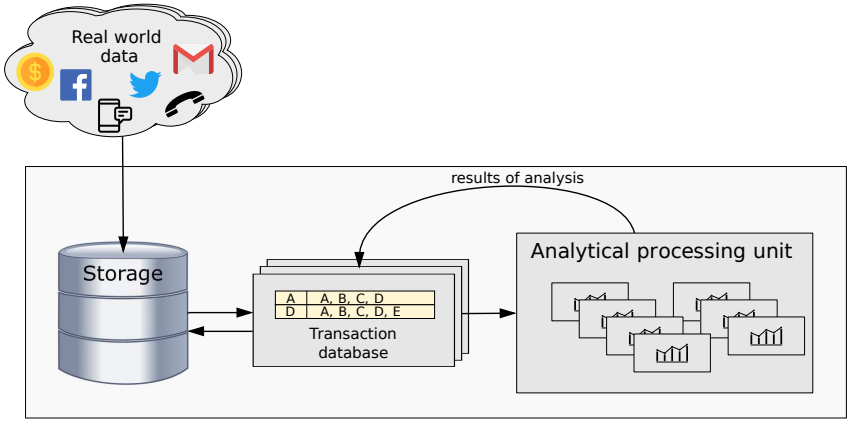
Analytical Operators

- $f : (TD, params) \rightarrow TD$
- Pattern mining and similarity search
  - Strong analytical tools
  - Pattern mining discover unknown
  - Similarity analysis looks for known



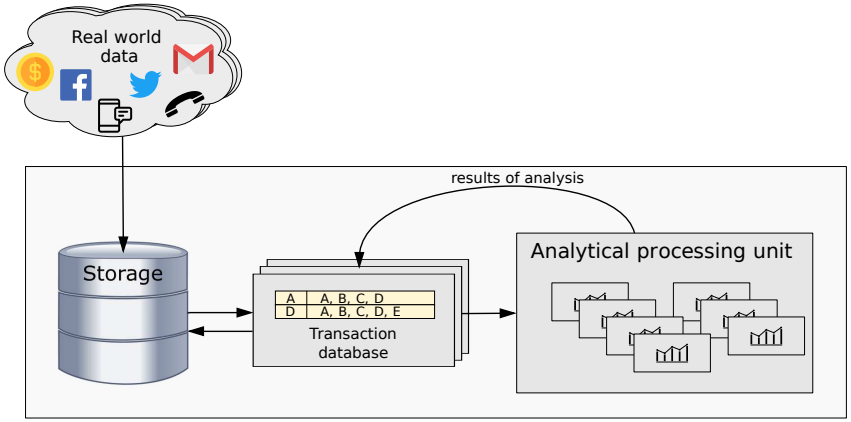
Analytical Operators

- $f : (TD, params) \rightarrow TD$
- Pattern mining and similarity search
  - Strong analytical tools
  - Pattern mining discover unknown
  - Similarity analysis looks for known
- Pattern mining
  - Frequent item-set mining
  - Sequence mining
  - Association rule mining



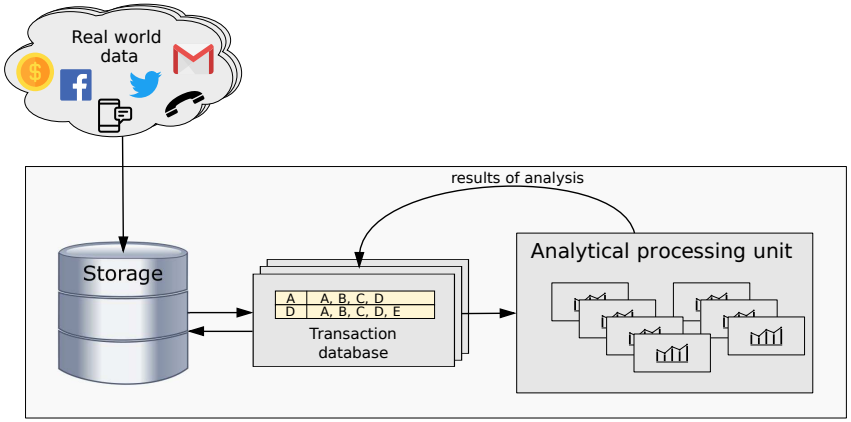
Analytical Operators

- $f : (TD, params) \rightarrow TD$
- Pattern mining and similarity search
  - Strong analytical tools
  - Pattern mining discover unknown
  - Similarity analysis looks for known
- Pattern mining
  - Frequent item-set mining
  - Sequence mining
  - Association rule mining
- Similarity in metric space
  - K-nn query
  - Range query
  - Similarity join



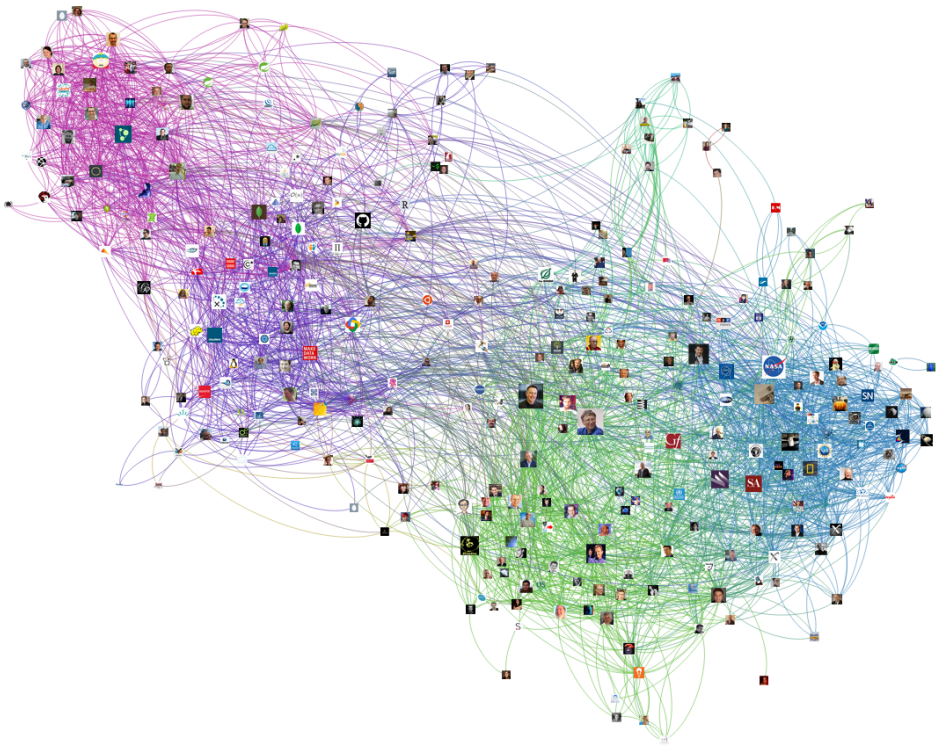
Analytical Operators

- $f : (TD, params) \rightarrow TD$
- Pattern mining and similarity search
  - Strong analytical tools
  - Pattern mining discover unknown
  - Similarity analysis looks for known
- Pattern mining
  - Frequent item-set mining
  - Sequence mining
  - Association rule mining
- Similarity in metric space
  - K-nn query
  - Range query
  - Similarity join
- Other operators?



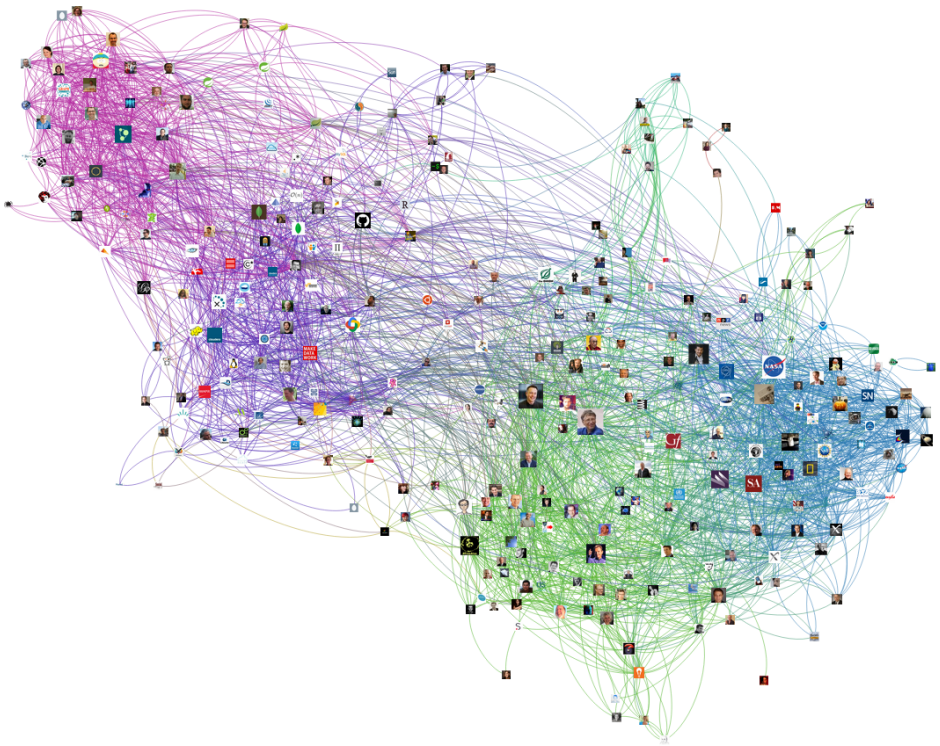
Use-cases

- Management of social network community



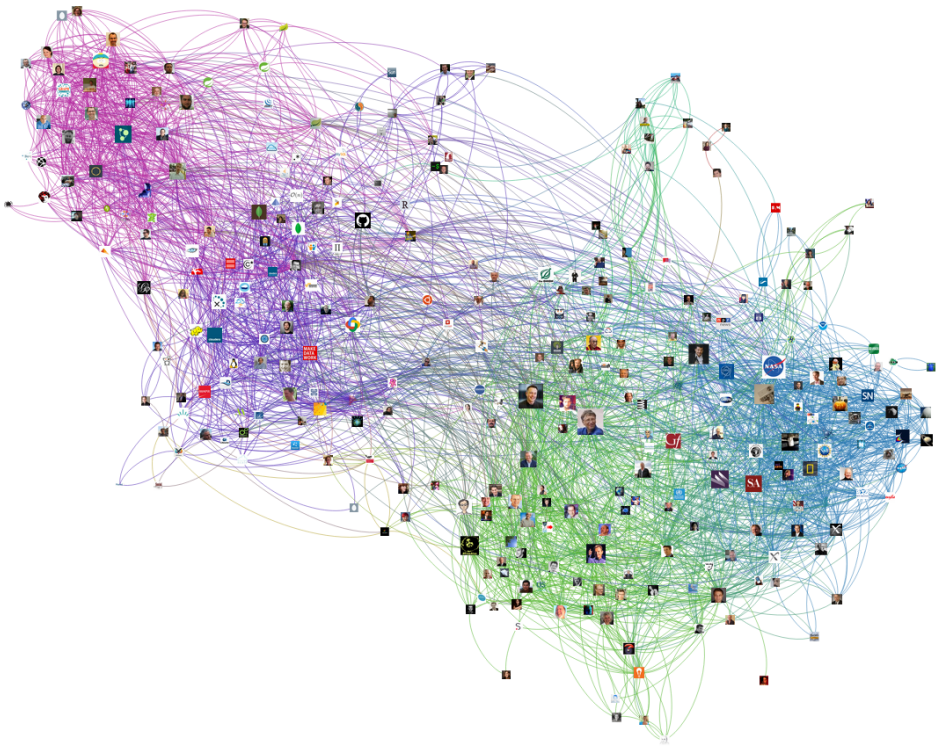
Use-cases

- Management of social network community
  - Group detection



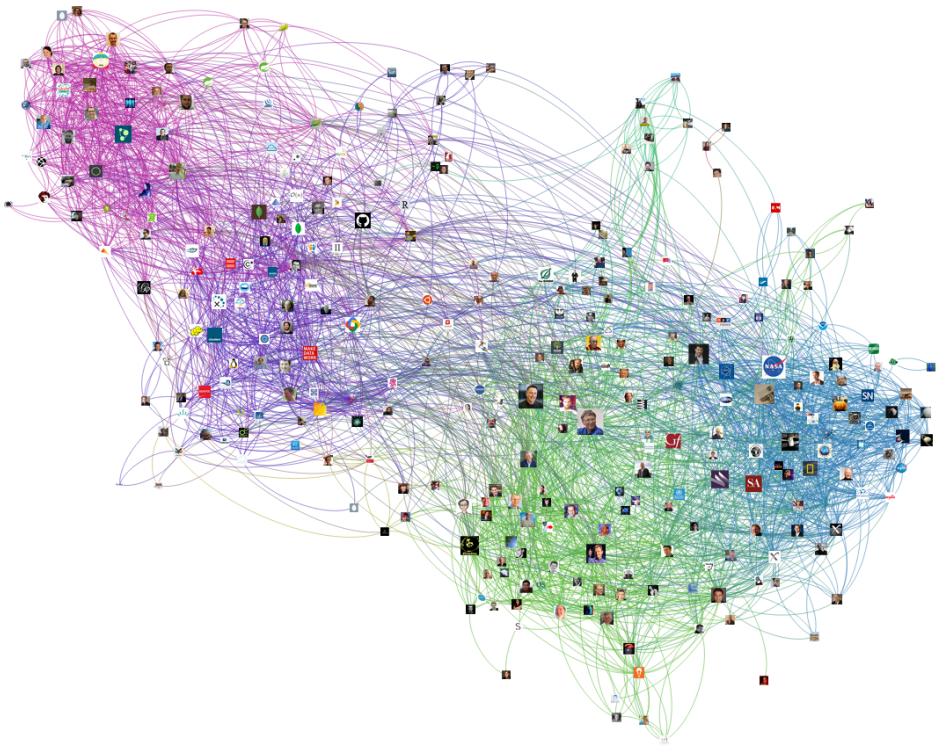
Use-cases

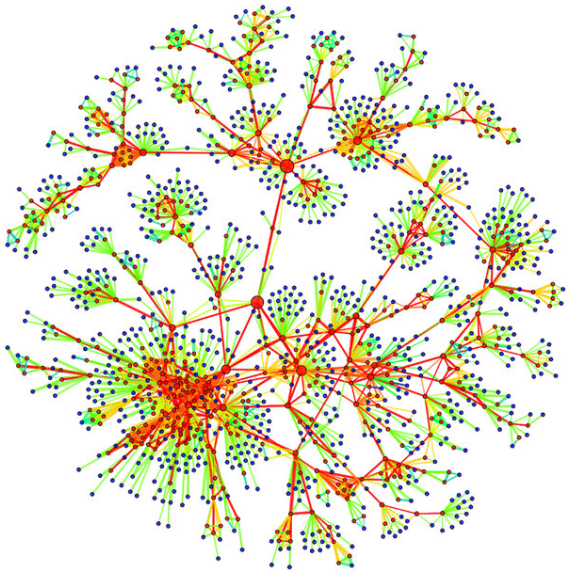
- Management of social network community
  - Group detection
  - Communication flows



Use-cases

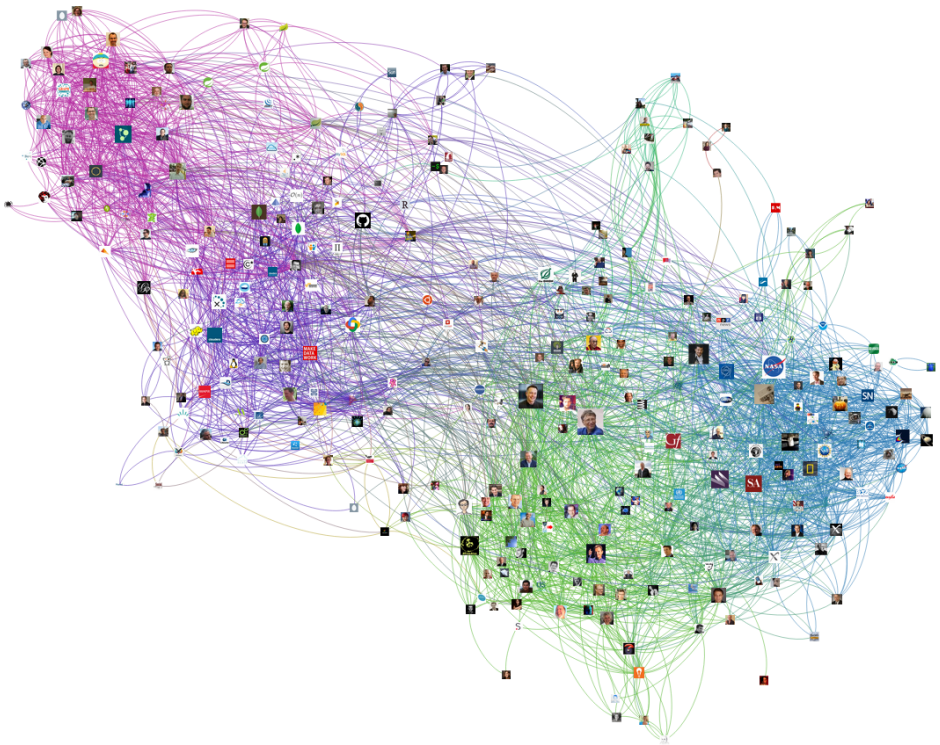
- Management of social network community
  - Group detection
  - Communication flows

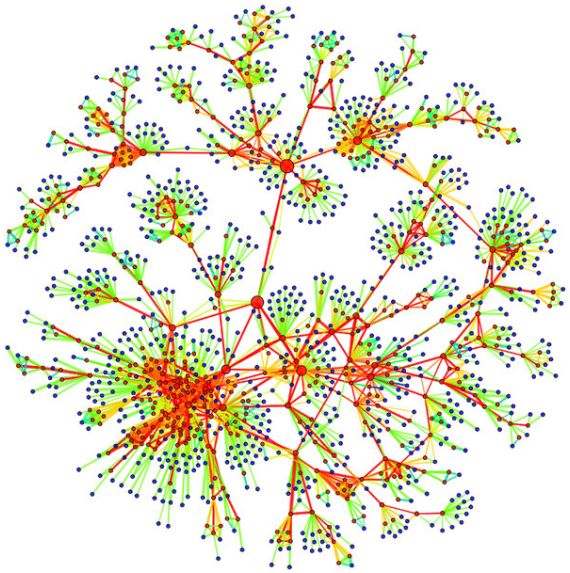




Use-cases

- Management of social network community
  - Group detection
  - Communication flows
  - Duplicate accounts detection





Experiments

- **Twitter Higg's boson dataset**
  - **Size**: 304 691 interactions on Twitter
- **Kosarak dataset**
  - **Size**: 990 000 click-streams through Hungarian news web
- **Community mining**:
  - 7 communities of size 12
  - 94 communities of size 11
- **Subsequence mining**:
  - Discovered 322 paths
  - 5 paths contained more than 4 nodes
  - Longest path has 16 nodes
- **Similarity search**:
  - Four nodes has most similar items inside community
  - One node has all ten outside of the community
  - Average amount of query nodes in range query results is 8.33 nodes

Summary

- What is goal?
  - Set of recommendations for task-oriented analysis
  - Universal system for analysis of data as proof-of-concept
- What we proposed?
  - Advanced Data Analysis by Mining and Searching System
  - Graph representation for capturing all the information
  - Transaction database as easily process-able format
  - Analytical operators: pattern mining, similarity search, etc.
- What it is for?
  - Analysis of communities
  - Analysis of sequences
  - Exploration by similarity searching
- What has been done?
  - Datasets: Twitter Higg's boson, Kosarak
  - Analysis of communities
  - Analysis of sequences
  - Similarity of neighbourhood of community members

Acknowledgements

This work has been supported by the Ministry of the Interior of the Czech Republic under the "Security Research for the Needs of the State Program 2015-2020," through the Project No. VI20172020096, "Complex Analysis and Visualization of Large-scale Heterogeneous Data."