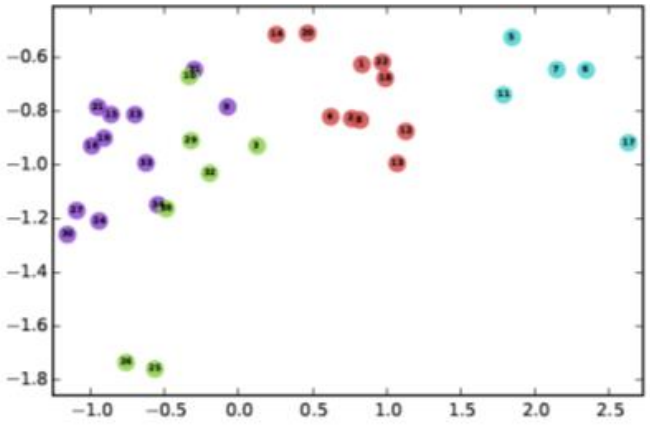
Cleora

SYNERISE

Barbara Rychalska

# Node embeddings

Input

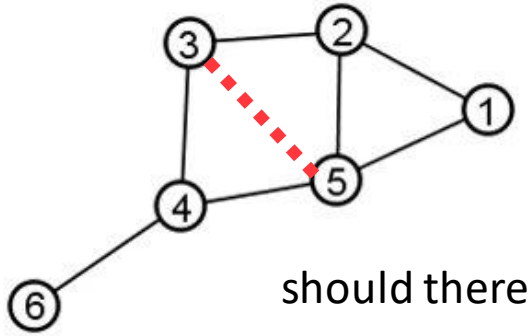



Output

- N-dimensional vector representation computed for each node
- The representations encode **node similarity**
  - 2 nodes are similar if they link to the same nodes
- They can also encode **node properties**
  - What is a node's degree?
  - Is it a leaf node?
  - Etc.
- They are usually reasonably small (of length 128 to 1024)

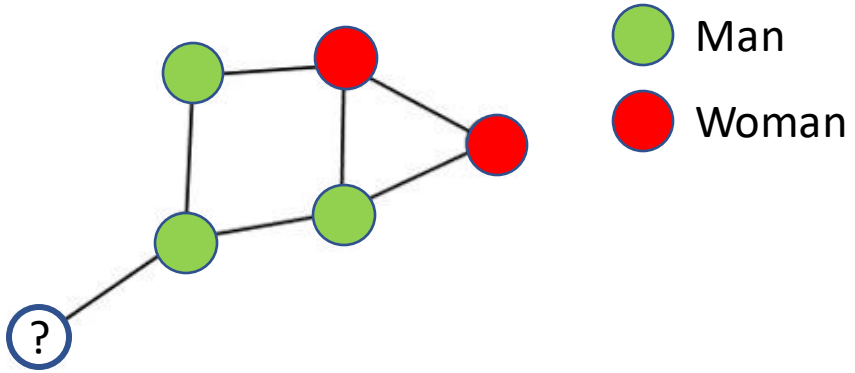
# Evaluating node embedding quality

Link Prediction:




should there be an edge?

Node Classification:



# Dataset scales

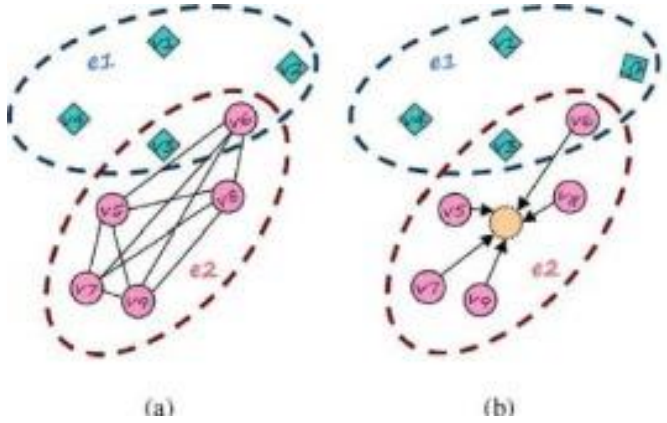
Hundreds of millions of nodes, billions of edges.

| Name | **Facebook** | **YouTube** | **RoadNet** | **LiveJournal** | **Twitter** |
|---|---|---|---|---|---|
| Nodes | 22,470 | 1,134,890 | 1,965,206 | 4,847,571 | 41,652,230 |
| Edges | 171,002 | 2,987,624 | 2,766,607 | 68,993,773 | 1,468,365,182 |
| Average Degree | 12 | 5 | 3 | 16 | |
| Density | $6 \times 10^{-4}$ | $4.7 \times 10^{-6}$ | $1.4 \times 10^{-6}$ | $3.4 \times 10^{-6}$ | |
| Classes | 4 | 47 | - | - | - |
| Directed | No | No | No | Yes | Yes |

Table 1: Dataset statistics.

# Cleora Aims

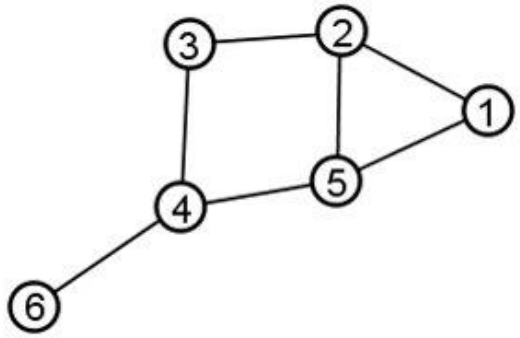
- Fast
- Scalable
- Simple
- Undirected edges (for now)
- CPU support
- Many modes of operation, i.e. hyperedge expansion




Clique Expansion Star Expansion

# Cleora Algorithm

1. Multiply M * $T_{i-1}$
2. Normalize $T_i$




d

|V|

| | |
|---|---|
| **0.23** | |
| 0.01 | |
| -0.2 | |
| -0.45 | |
| 0.06 | |
| 0.13 | |

$T_{i-1}$

M

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | | **1/2** | | | **1/2** | |
| 2 | 1/3 | | 1/3 | | 1/3 | |
| 3 | | 1/2 | | 1/2 | | |
| 4 | | | 1/3 | | 1/3 | 1/3 |
| 5 | 1/3 | 1/3 | | 1/3 | | |
| 6 | | | | 1 | | |

| | | | |
|---|---|---|---|
| **0.035** | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

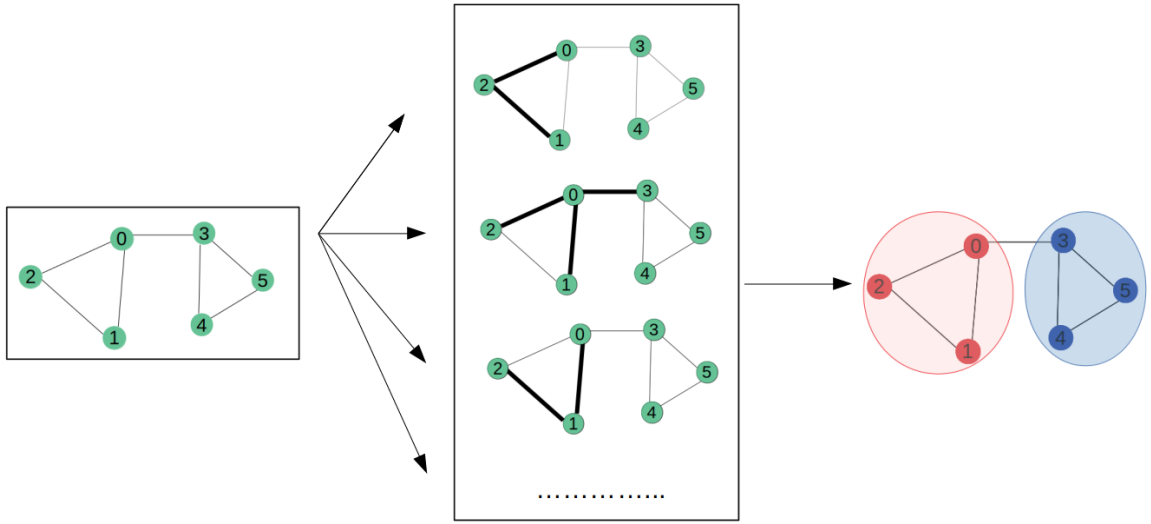
$T_i$

# Why does it work?

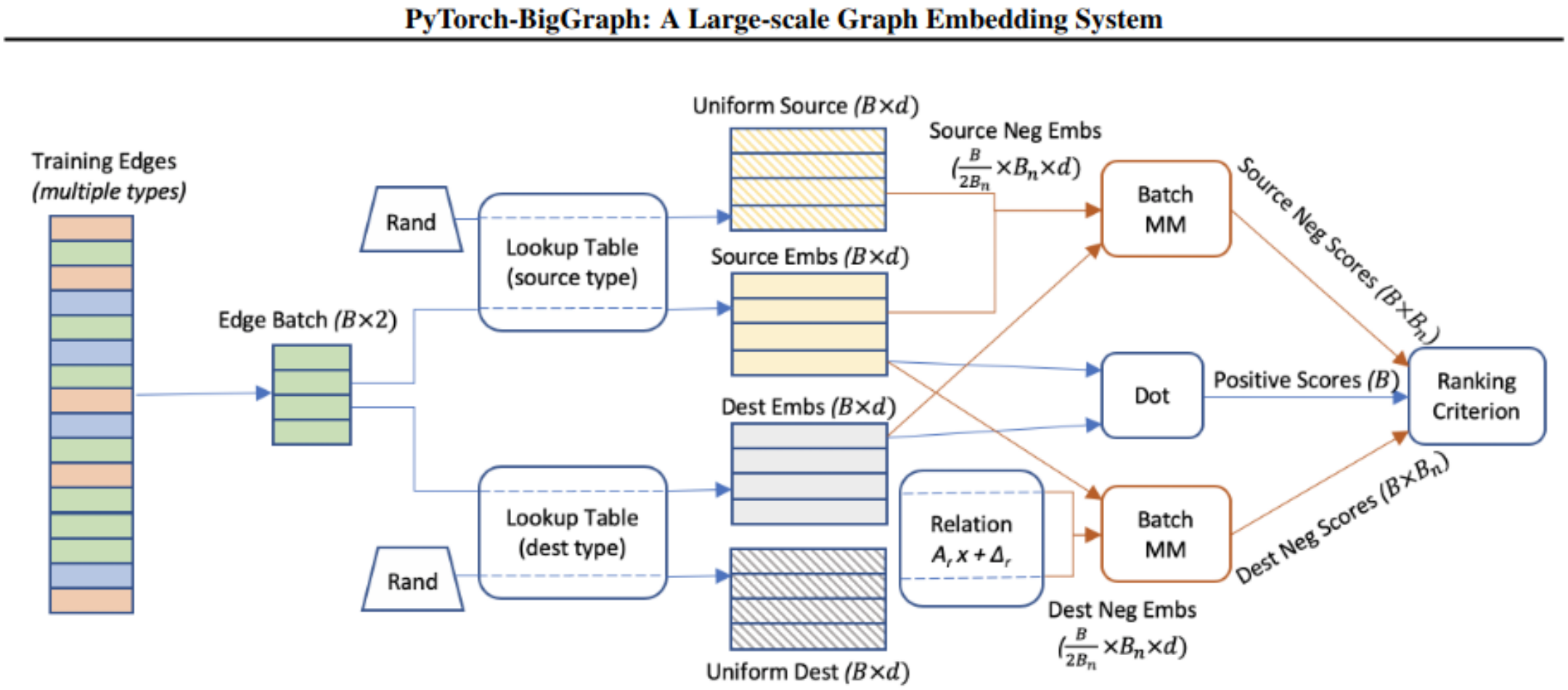
- Graph Convolutional Networks
  - The graph convolution operation
  - Proven that it works only partially
    - http://proceedings.mlr.press/v97/wu19e.html
  - Cleora uses ~the most important part

$$H^{[i+1]} = \sigma(W^{[i]} H^{[i]} A^{*})$$

Equation 3— Forward Pass in Graph Convolutional Networks

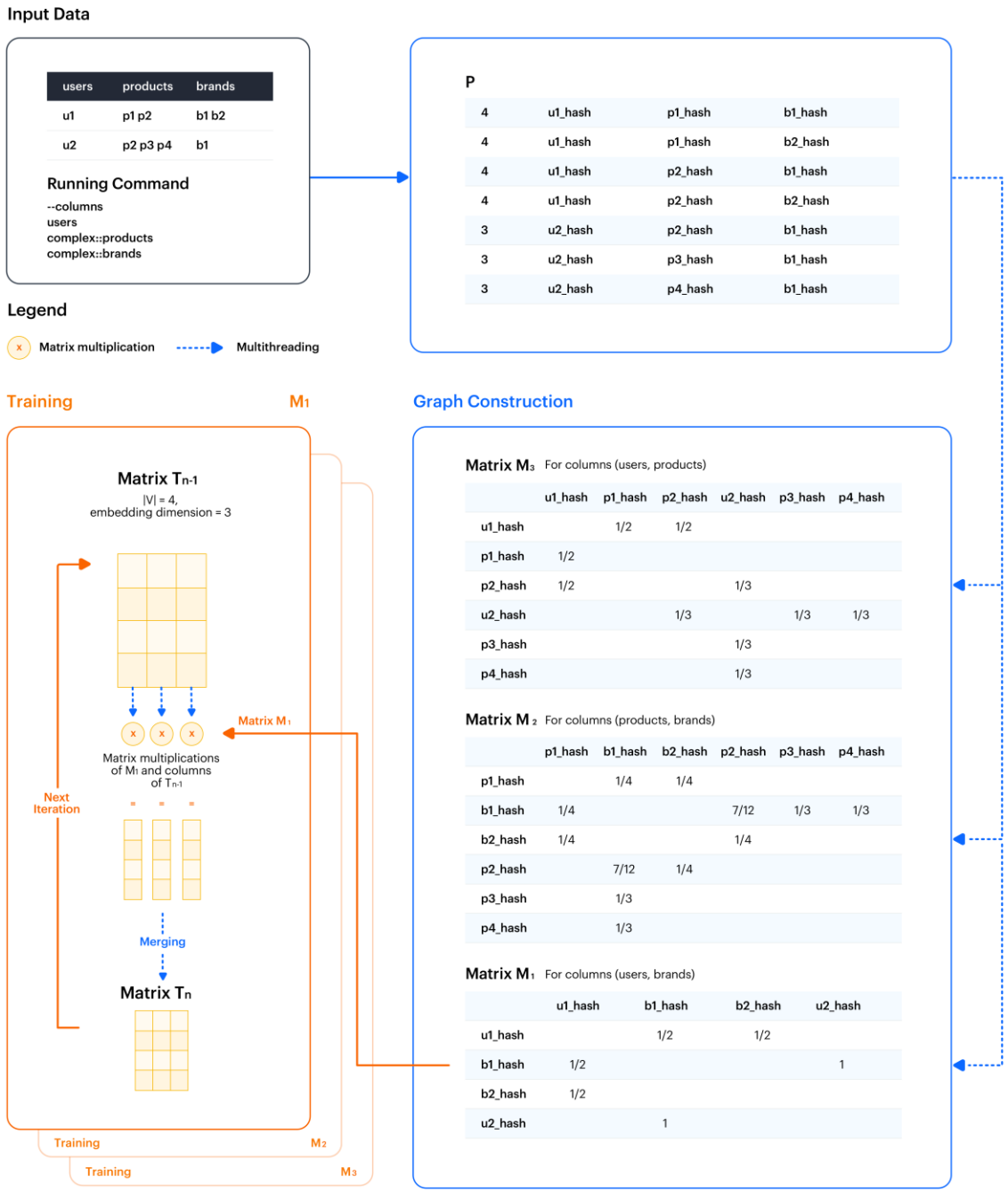


# Complexity of competitors



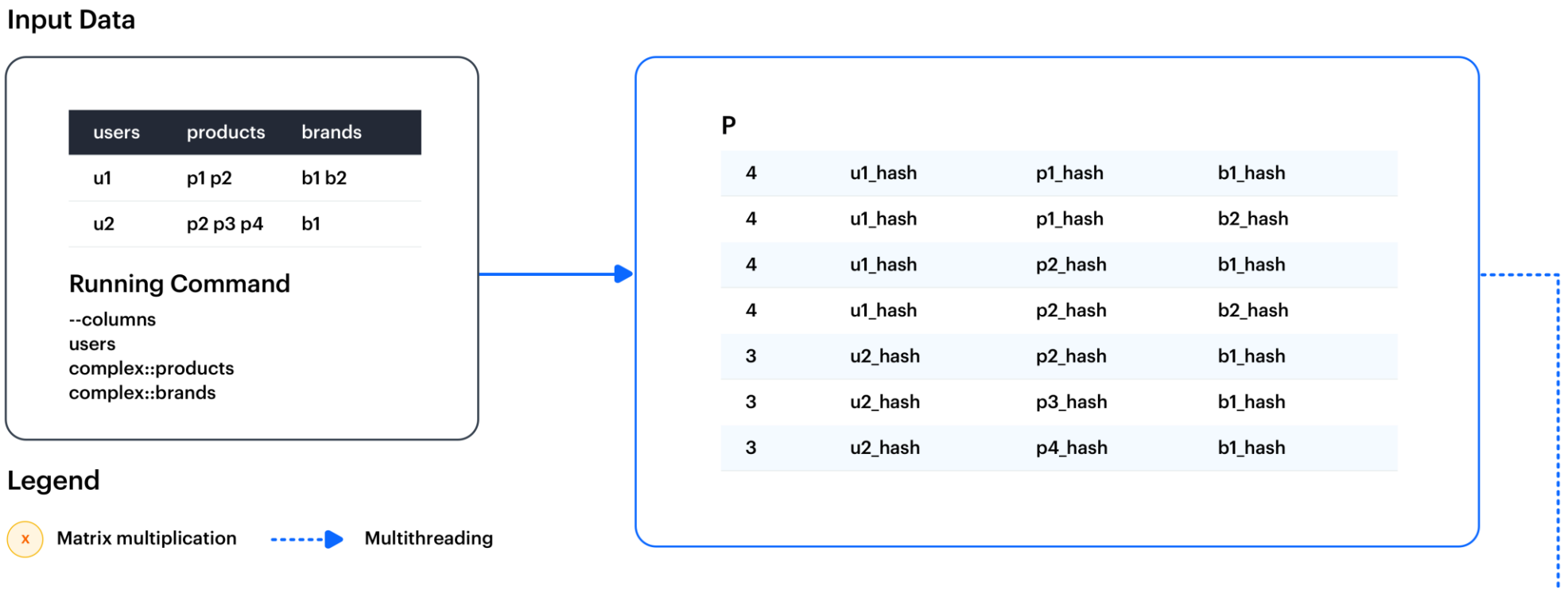
# Cleora implementation

- Implemented in Rust
- Uses custom-made SparseMatrix struct
  - Only (x,y) coordinates of nonzero values are stored
- Uses fast non-cryptographic hash functions
- Uses CPU fast access (data locality)
- Parallel collection for matrix multiplication
- Big graphs calculation can be supported by memory-mapped files



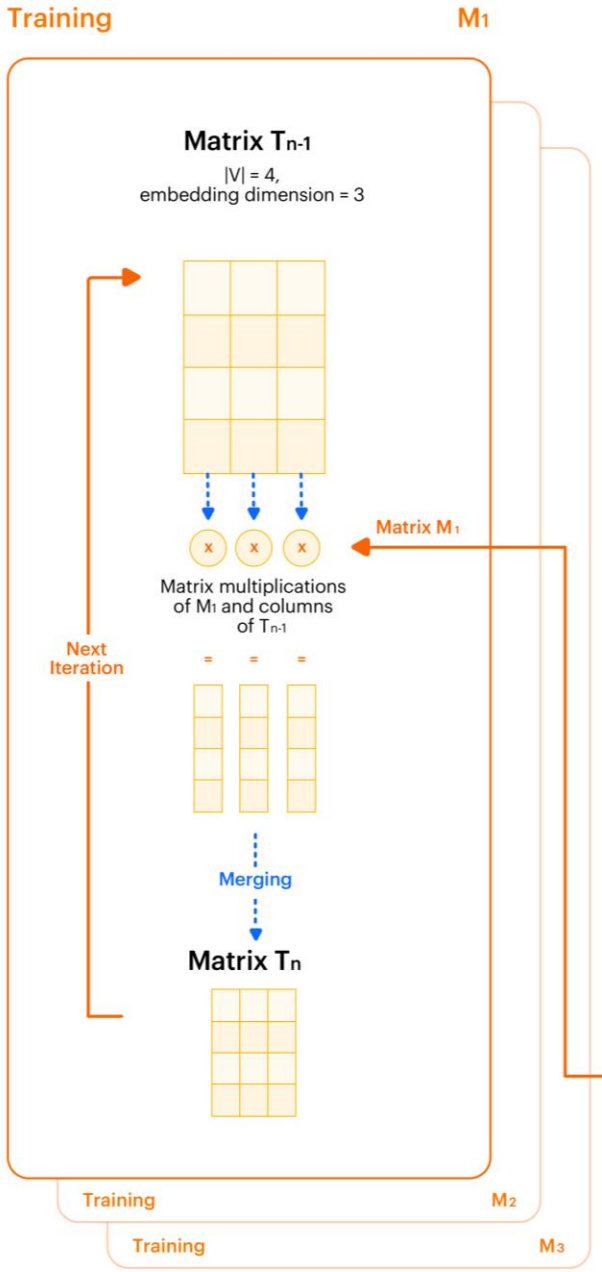
# Cleora implementation

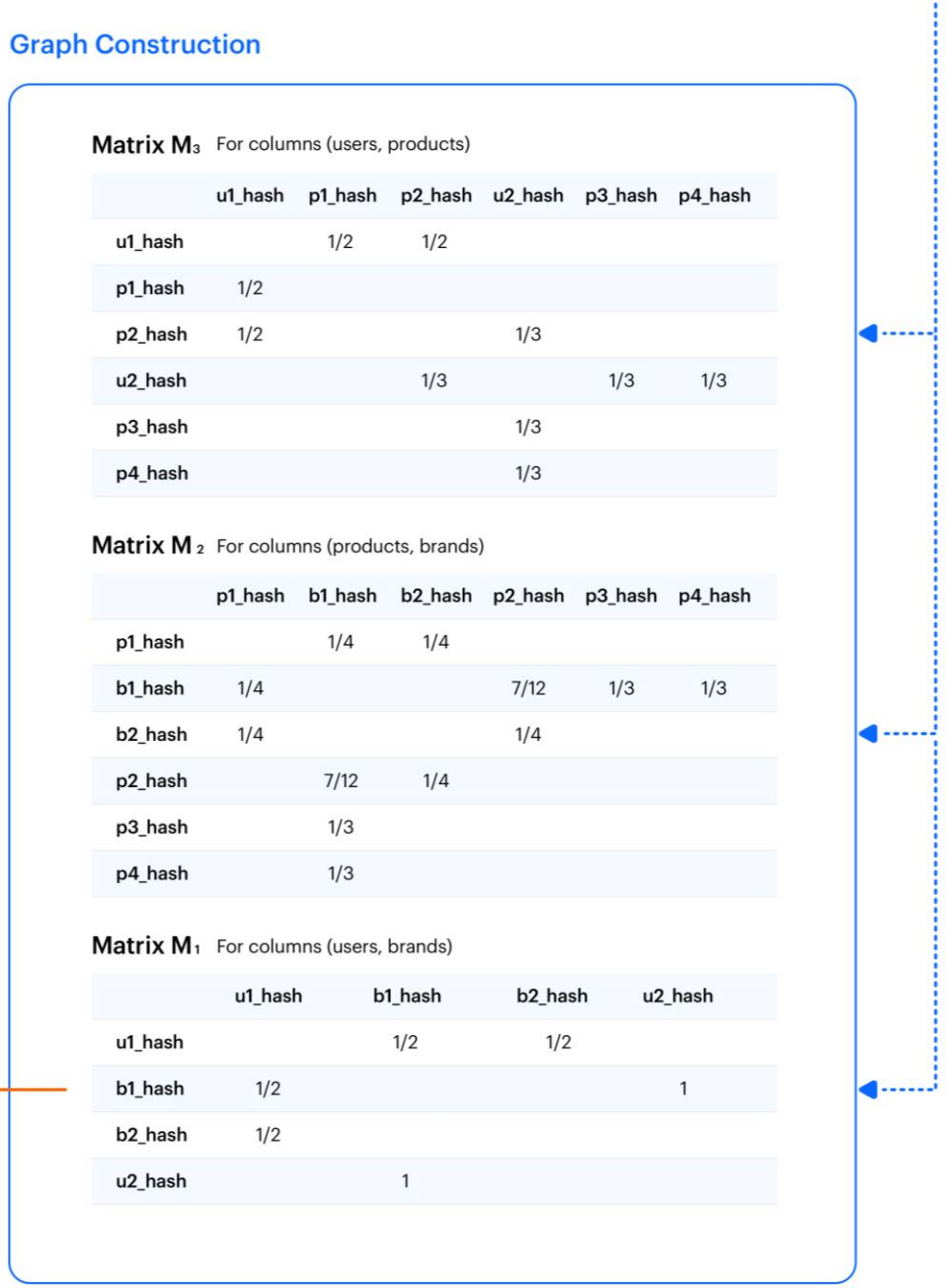
Input format & helper matrix P



# Cleora implementation

Graph construction & training




**Matrix $M_3$** For columns (users, products)

| | u1_hash | p1_hash | p2_hash | u2_hash | p3_hash | p4_hash |
|---|---|---|---|---|---|---|
| u1_hash | | 1/2 | 1/2 | | | |
| p1_hash | 1/2 | | | | | |
| p2_hash | 1/2 | | | 1/3 | | |
| u2_hash | | | 1/3 | | 1/3 | 1/3 |
| p3_hash | | | | 1/3 | | |
| p4_hash | | | | 1/3 | | |

**Matrix $M_2$** For columns (products, brands)

| | p1_hash | b1_hash | b2_hash | p2_hash | p3_hash | p4_hash |
|---|---|---|---|---|---|---|
| p1_hash | | 1/4 | 1/4 | | | |
| b1_hash | 1/4 | | | 7/12 | 1/3 | 1/3 |
| b2_hash | 1/4 | | | 1/4 | | |
| p2_hash | | 7/12 | 1/4 | | | |
| p3_hash | | 1/3 | | | | |
| p4_hash | | 1/3 | | | | |

**Matrix $M_1$** For columns (users, brands)

| | u1_hash | b1_hash | b2_hash | u2_hash |
|---|---|---|---|---|
| u1_hash | | 1/2 | 1/2 | |
| b1_hash | 1/2 | | | 1 |
| b2_hash | 1/2 | | | |
| u2_hash | | 1 | | |

# Cleora - properties

- Speed

| Algorithm | Facebook | YouTube | RoadNet | LiveJournal | Twitter |
|---|---|---|---|---|---|
| **Total embedding time** | | | | | |
| Cleora | 00:00:43 h | 00:12:07 h | 00:24:15 h | 01:35:40 h | 25:34:18 h |
| PBG | 00:04.33 h | 00:54:35 h | 00:37:41 h | | -* |
| Deepwalk | 00:36:51 h | 28:33:52 h | 53:29:13 h | *timeout* | *timeout* |
| **Training time** | | | | | |
| Cleora | 00:00:25 h | 00:11:46 h | 00:04:14 h | 01:31:42 h | 17:14:01 h |
| PBG | 00:02:03 h | 00:24:15 h | 00:31:11 h | 07:10:00 h | -* |

- Accuracy

| **Algorithm** | **Facebook** | | **YouTube** | | **RoadNet** | | **LiveJournal** | | **Twitter** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MRR | HR@10 | MRR | HR@10 | MRR | HR@10 | MRR | HR@10 | MRR | HR@10 |
| Scalable methods | | | | | | | | | | |
| Cleora | 0.0724 | 0.1804 | 0.0471 | 0.0618 | 0.9243 | 0.9429 | 0.6079 | 0.6665 | 0.0355 | 0.076 |
| PBG [1] | 0.0817* | 0.2133* | 0.0321* | 0.0640* | 0.8717* | 0.9106* | 0.5549* | 0.6770* | -** | -** |
| GOSH [2] | 0.0924* | 0.2319* | 0.0280* | 0.0590* | 0.8756* | 0.8977* | 0.2242* | 0.4012* | -** | -** |
| Non-scalable methods | | | | | | | | | | |
| Deepwalk [7] | 0.0803* | 0.1451* | 0.1045* | 0.1805* | 0.9626* | 0.9715* | *timeout* | *timeout* | *timeout* | *timeout* |
| LINE [23] | 0.0749* | 0.1923* | 0.1064* | 0.1813* | 0.9628* | 0.9833* | 0.5663* | 0.6670* | -** | -** |

Dataset stats

| Name | Facebook | YouTube | RoadNet | LiveJournal | Twitter |
|---|---|---|---|---|---|
| Nodes | 22,470 | 1,134,890 | 1,965,206 | 4,847,571 | 41,652,230 |
| Edges | 171,002 | 2,987,624 | 2,766,607 | 68,993,773 | 1,468,365,182 |
| Average Degree | 12 | 5 | 3 | 16 | |
| Density | $6 \times 10^{-4}$ | $4.7\times10^{-6}$ | $1.4\times10^{-6}$ | $3.4\times10^{-6}$ | |
| Classes | 4 | 47 | - | - | - |
| Directed | No | No | No | Yes | Yes |

Table 1: Dataset statistics.

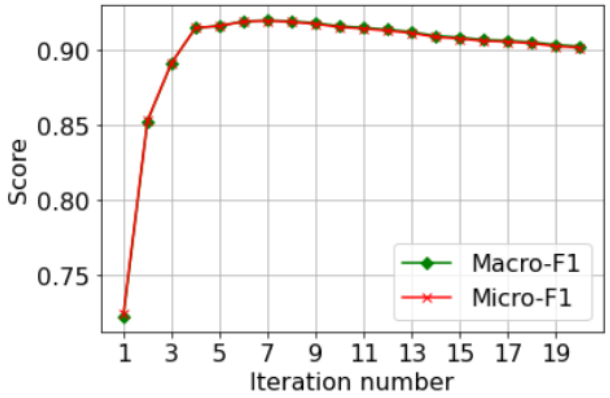
| **Algorithm** | **Facebook** | | **YouTube** | |
|---|---|---|---|---|
| | Micro-F1 | Macro-F1 | Micro-F1 | Macro-F1 |
| Scalable methods | | | | |
| Cleora | 0.9190 | 0.9191 | 0.3859 | 0.3077 |
| PBG | 0.9258 | 0.9262 | 0.3567* | 0.2459* |
| GOSH | 0.8312* | 0.8305* | 0.3166* | 0.2245* |
| Non-scalable methods | | | | |
| Deepwalk | 0.9349* | 0.9354* | 0.3166* | 0.2245* |
| LINE | 0.9442* | 0.9446* | 0.4008* | 0.3338* |

Table 4: Classification performance results. * - results with statistically significant differences to Cleora according to the Wilcoxon two-sided paired test (p-value lower than 0.05).

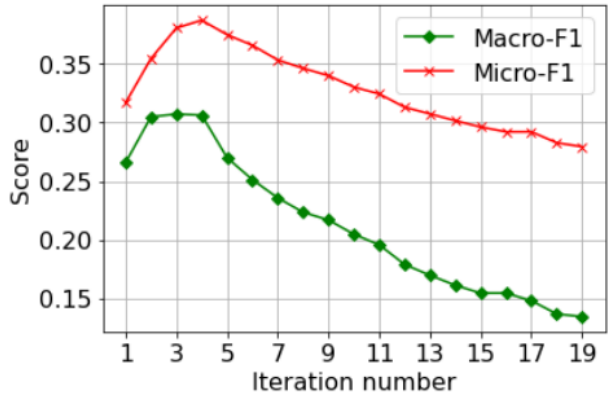
SYNERISE

# Cleora - properties

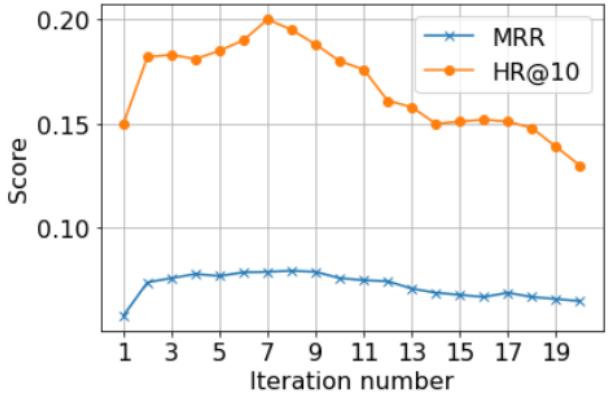
- Iteration number defines performance on link prediction & classification tasks



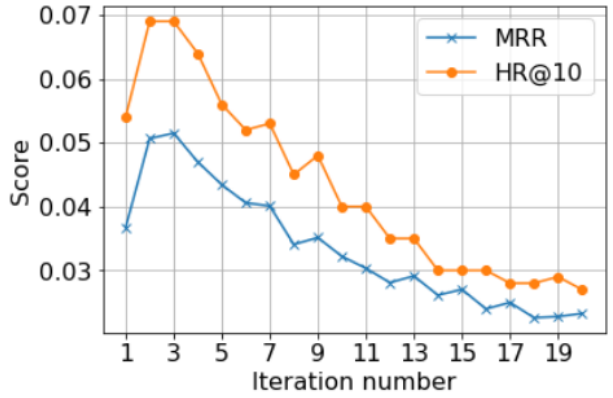

(a) Facebook Dataset - Classification Task.




(b) Youtube Dataset - Classification Task




(c) Facebook Dataset - Link Prediction Task.




(d) Youtube Dataset - Link Prediction Task.

Figure 2: The influence of iteration number on embedding quality.

# Cleora - properties

- Inference of new node embedding = weighted averaging of neighbors' embeddings
- Embedding large graphs which don't fit into RAM:
  1. Chunk graph into N parts
  2. Compute embeddings for each chunk
  3. Perform weighted averaging of per-node embeddings from each chunk
- Iteration number changes behavior: complement vs substitute prediction

**Substitute**: Items bought in place of another, e.g. Samsung TV & LG TV

**Complement**: Items bought together: e.g. iPod and headphones

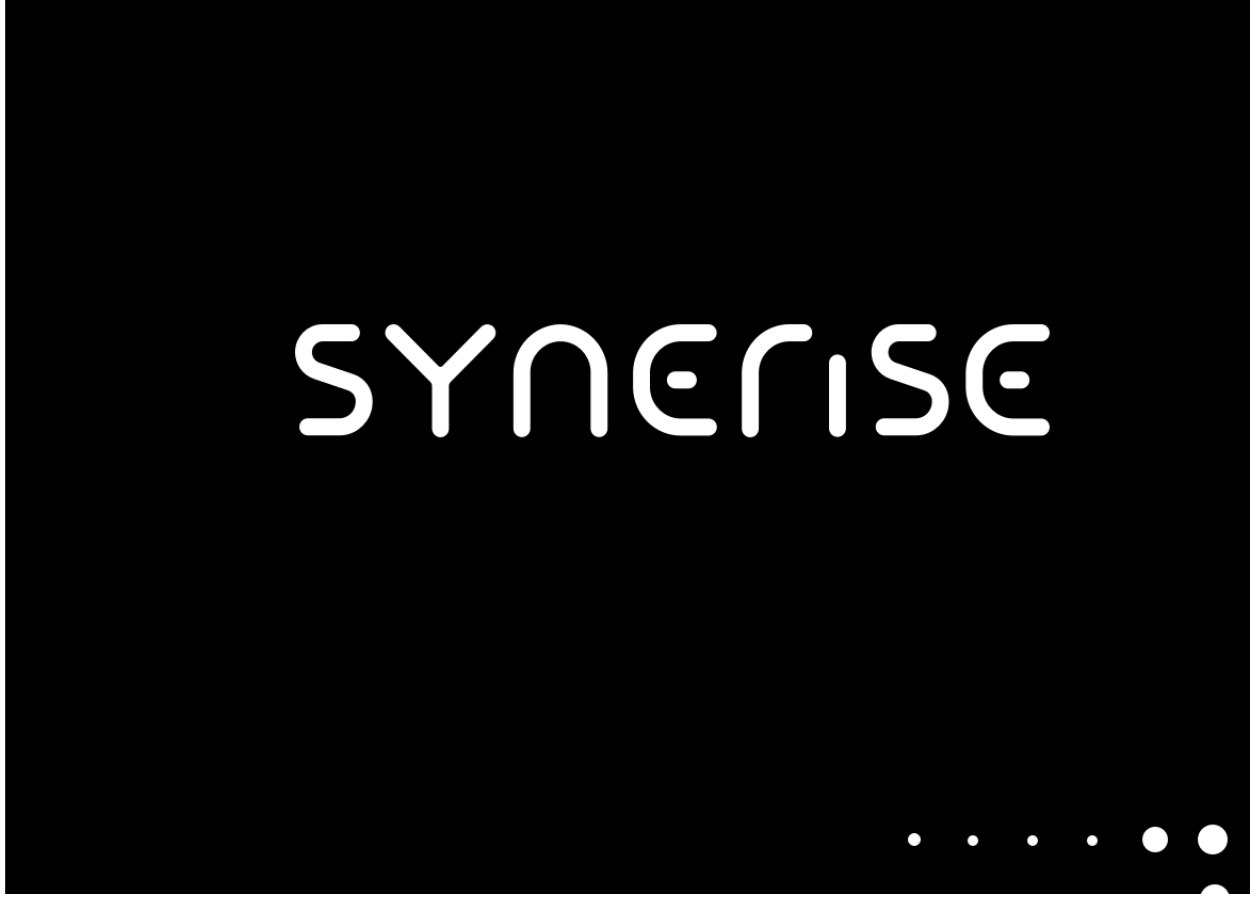
| | SOUP RAMEN NOODLES/RAMEN CUPS 3 OZ | |
|---|---|---|
| | 1 iteration | 4 iterations |
| 1. | AUTOMOTIVE PRODUCTS 4 CT | SOUP RAMEN NOODLES/RAMEN CUPS 3 OZ |
| 2. | PROCESSED DIPS 15.5 OZ | SOUP RAMEN NOODLES/RAMEN CUPS 3 OZ |
| 3. | SOUP RAMEN NOODLES/RAMEN CUPS 3 OZ | SOUP RAMEN NOODLES/RAMEN CUPS 3 OZ |
| 4. | J-HOOKS JHOOK - HOUSEWARE | SOUP RAMEN NOODLES/RAMEN CUPS 3 OZ |
| 5. | PACKAGED CANDY BAGS-CHOCOLATE 11 OZ | SOUP RAMEN NOODLES/RAMEN CUPS 3 OZ |
| | **BAKED BREAD/BUNS/ROLLS MAINSTREAM WHITE BREAD 20 OZ** | |
| | 1 iteration | 4 iterations |
| 1. | BAKED BREAD/BUNS/ROLLS DINNER ROLLS 11 OZ | SMOKED MEATS MARINATED |
| 2. | CHIPS&SNACKS MISC 3.5 OZ | PICKLE/RELISH/PKLD VEG PICKLES |
| 3. | SPRING/SUMMER SEASONAL SALLY HANSEN | PNT BTR/JELLY/JAMS JELLY |
| 4. | DRY NOODLES/PASTA SPAGHETTI DRY 16 OZ | COLD CEREAL KIDS CEREAL |
| 5. | BEERS/ALES BEERALEMALT LIQUORS 40 OZ | BREAKFAST SAUSAGE/SANDWICHES PATTIES |
| | **BREAD BREAD:ITALIAN/FRENCH** | |
| | 1 iteration | 4 iterations |
| 1. | LUNCHMEAT PEPPERONI/SALAMI 3 OZ | REFRIGERATED DOUGH PRODUCTS ROLLS |
| 2. | CANDY BAGS-NON CHOCOLATE 4.25 OZ | SEAFOOD - FROZEN SEAFOOD-FRZ-RW-ALL |
| 3. | GREETING CARDS/WRAP/PARTY SPLY PARTY | BAKED SWEET GOODS SNACK CAKE - PACK 5.7 OZ |
| 4. | VALENTINE VALENTINE GIFTWARE/DECOR 5 CT | PIES PIES: PUMPKIN/CUSTARD |
| 5. | CANDY - CHECKLANE CANDY BARS (SINGLES) | LUNCHMEAT HAM 9 OZ |

Table 7: Examples of complement vs substitute prediction.

# Cleora on Github

https://github.com/Synerise/cleora

- Released on MIT license – can use commercially
- Easy to run: just type `./cleora` on Linux-based systems; no extra package installation required


Thank you!