DCA, CIKM'16

Dmitry Dremov, Julia Ivkina, "ML Trainings"
3rd place

“ML Trainings”

- Moscow community
- Discussing competitions, experience sharing
- DCA: 3, 4, 7, 10 @ Top-10

Plan

- Train set transformation
- Filtration
- Features
- Edge probability prediction
- Closure

Train set transformation

- Find components
- Add random components up to 96000 nodes
- Add 443 random nodes from distinct components
- Two folds with 197352 nodes and 412945 edges total

Filtration

- kNN with visited URLs sparse matrix
- Shared URLs
- Train set edges ~4000000
- Recall ~53%
- Precision ~5%

kNN

- URLs sparse matrix
- Every visit “A/B/C/D” : {“A”, “A/B”, “A/B/C”, “A/B/C/D”}
- TF-IDF, row-wise L_2 normalisation
- L_2 distance
- 15 nearest neighbour

Shared URLs

- URLs visited by less than 40 devices.
- These devices are frequently linked
- Count shared URLs for every edge

Features

- Visited URLs-based
- Time-based

Visited URLs

- Sparse visited URLs matrix
- L_1, L_2, L_{max}, KLD, cosine
- Edge AB: (A_{URLs}, B_{URLs})
- SVD, 40 components

Time based

- Intraday usage
- Consecutive visits interval
- Time window intersection

Edge probability

- OOF base models usage
- Base RF and SVM models
- Meta XGBoost model

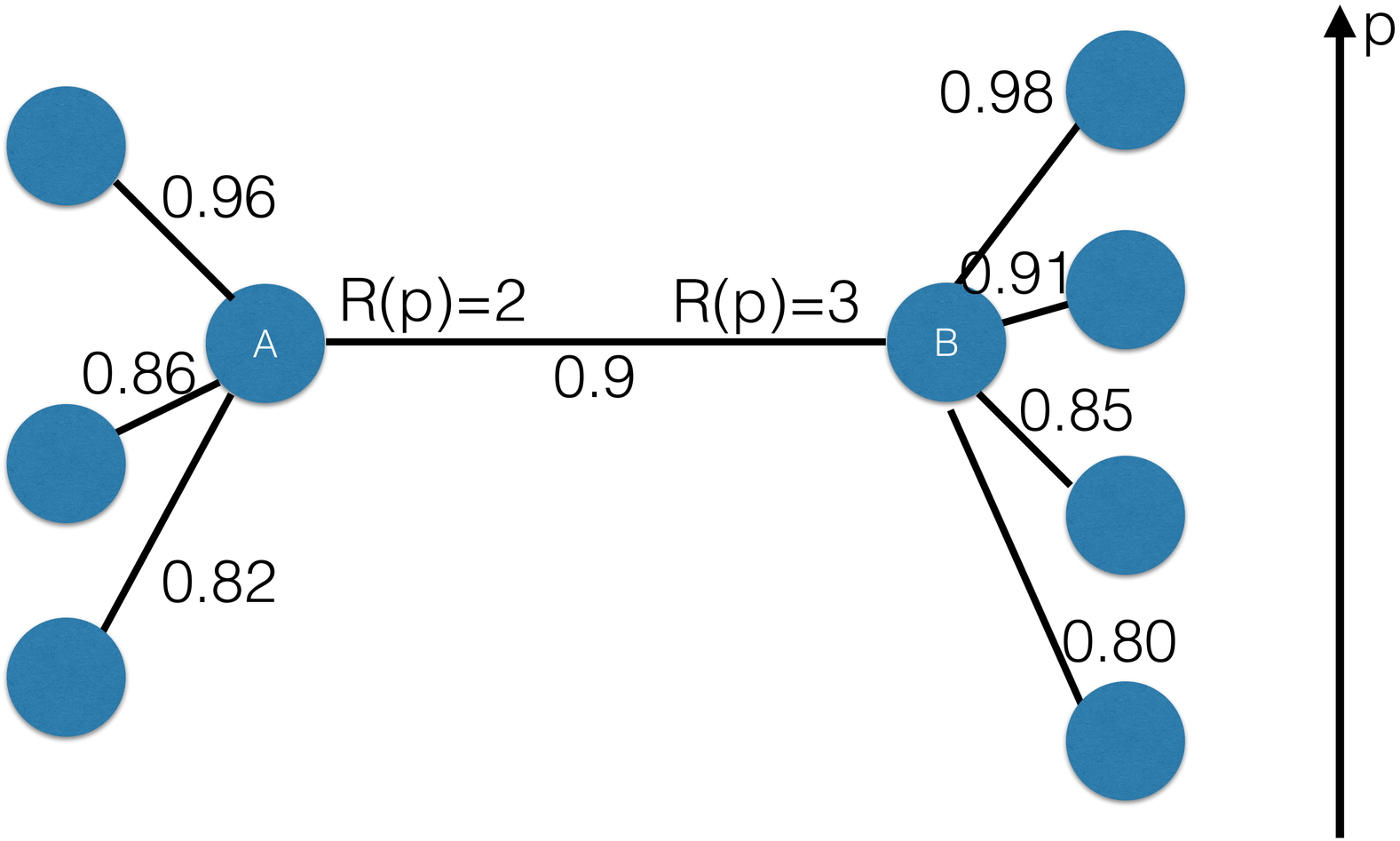
Base models

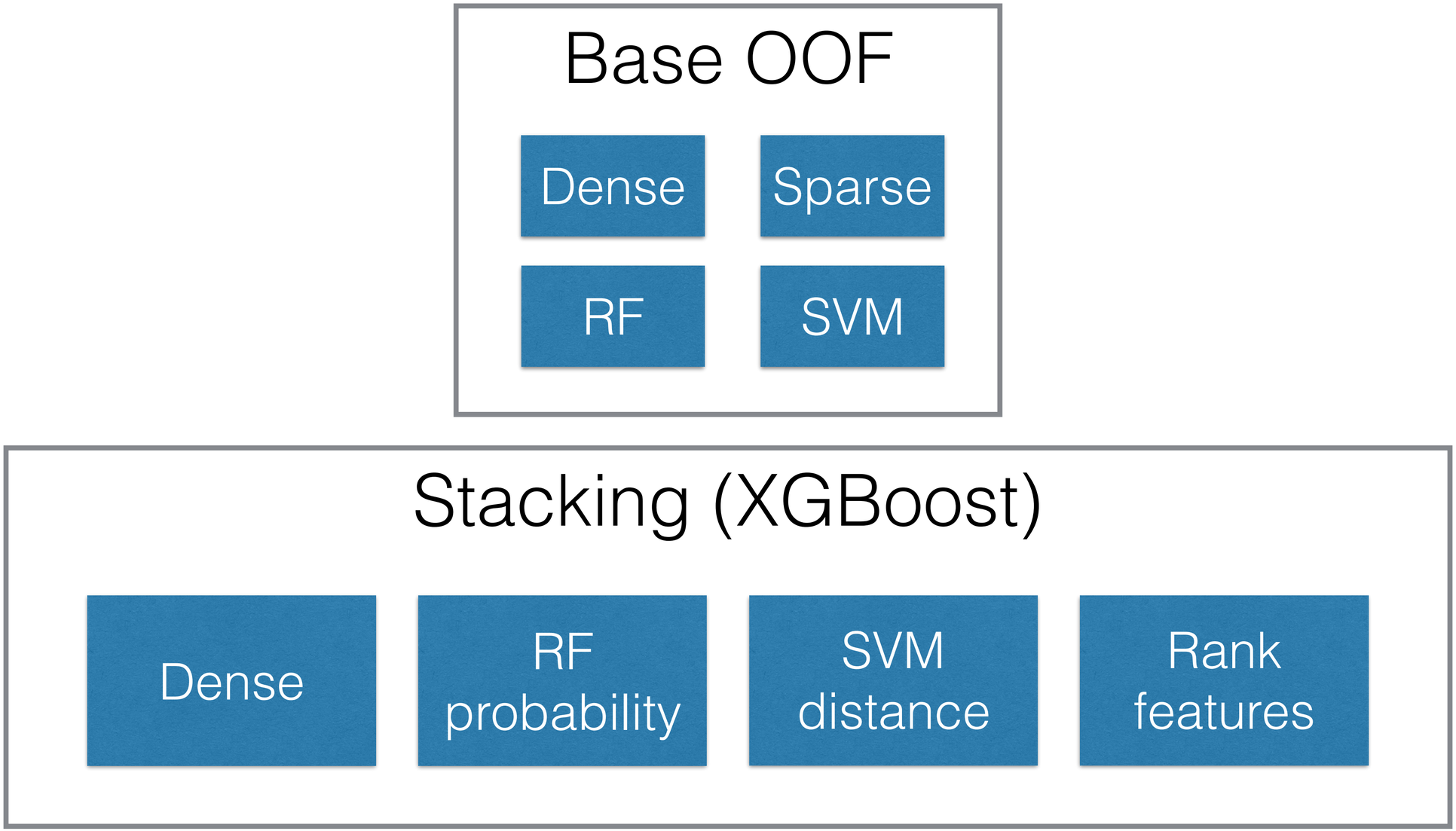
- SVM on sparse URLs matrix
- Separate model for every neighbour number, high regularisation
- Random Forest on dense features

Meta features

- Dense features
- Random Forest probability
- SVM boundary distance
- “Rank” features

Rank features

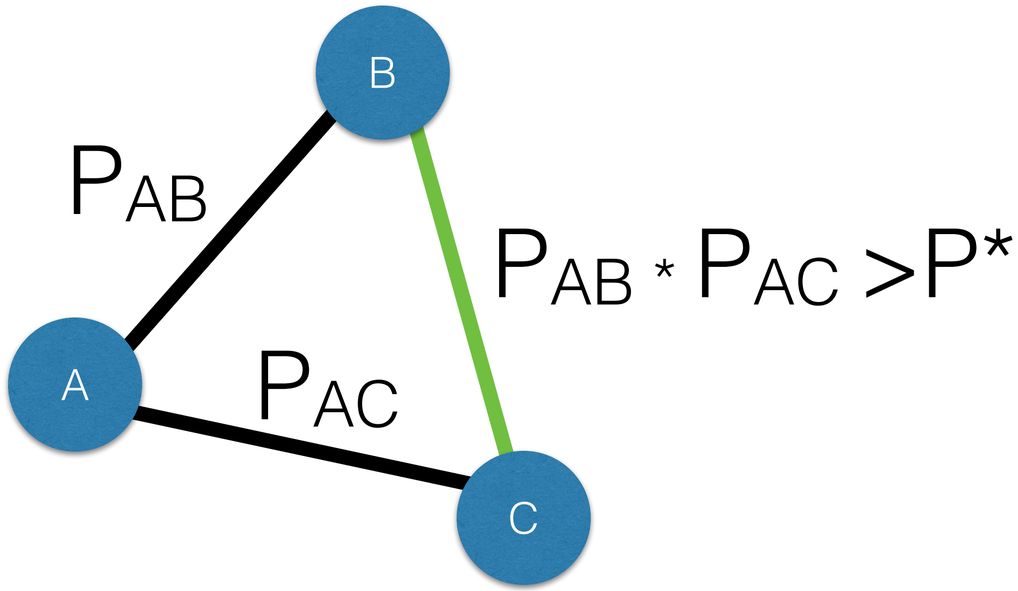



Base OOF
Dense
Sparse
RF
SVM
Stacking (XGBoost)
Dense
RF probability
SVM distance
Rank features

Closure

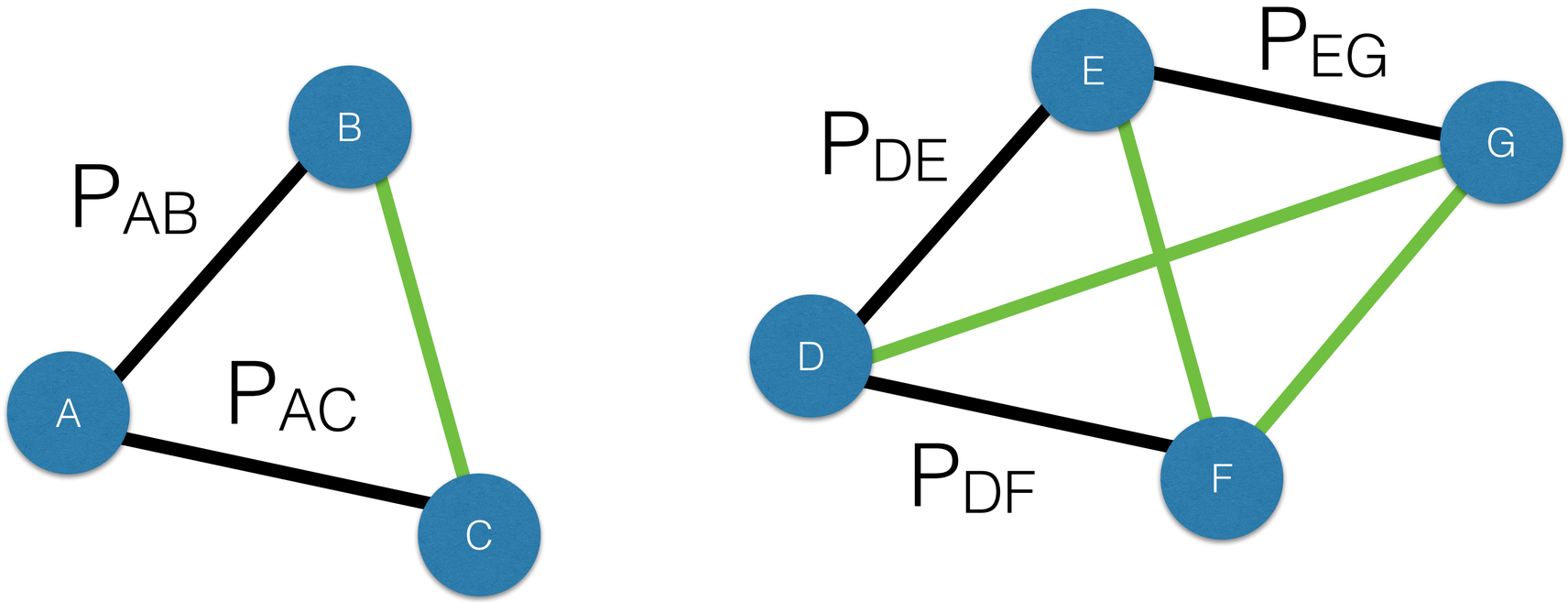
- Transitive closure
- Fully connected closure

Transitive closure



$P^* \sim 0.8$

Fully connected closure



$P^* \sim 0.9$

Resulted edges set

- Top edges by model 97000
- Transitive closure +12000 edges
- Fully connected closure +4000

Calculations

- 12 cores, 64 GB RAM
- ~4 hours total work time

Analysis

- 3rd place with F1* ~ 0.4137
- Strong model
- Good filtration
- Good features

Thank you!