Mining and Processing Biomedical Data

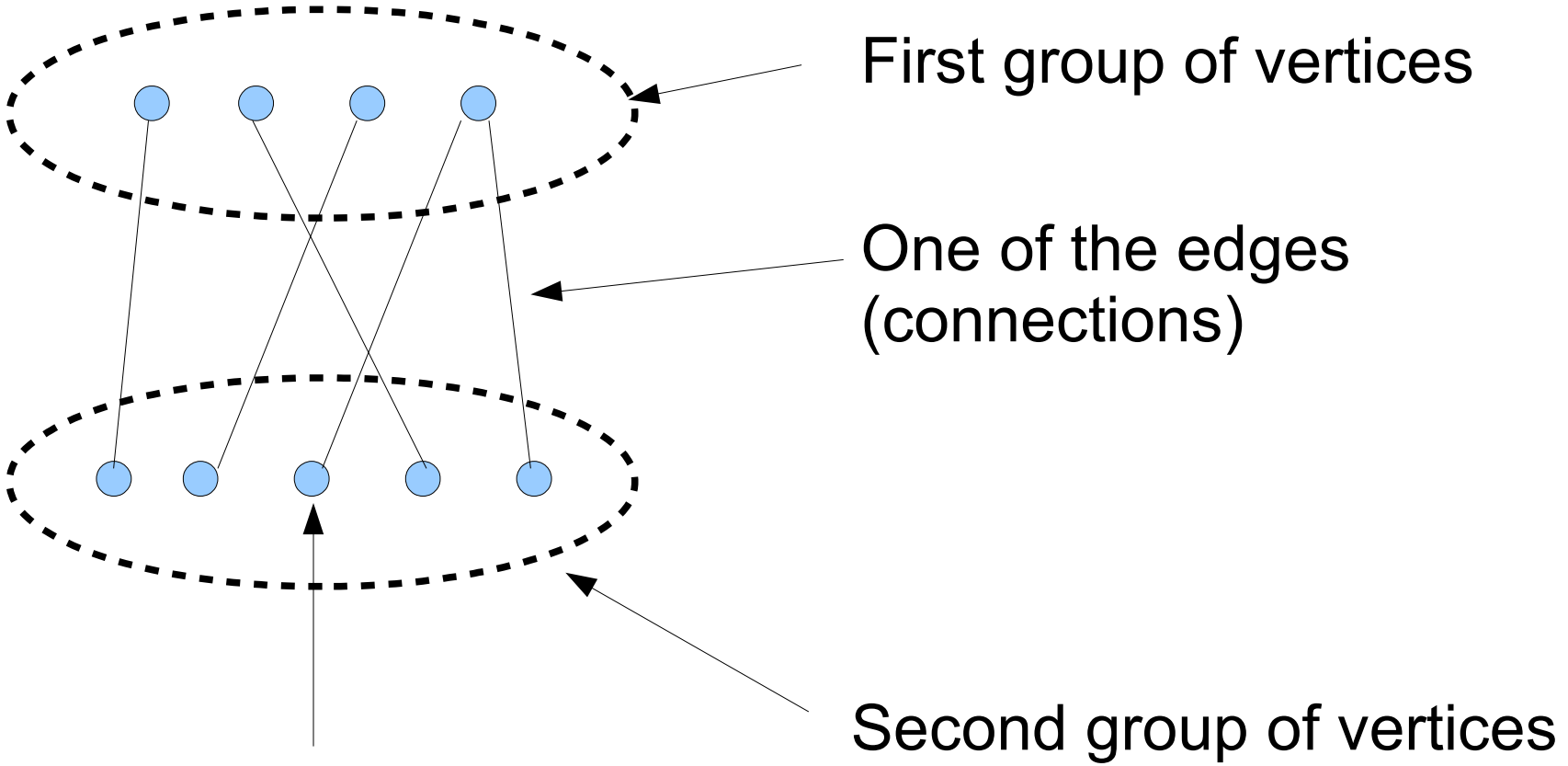
Dr. rer. nat. Krisztian Buza
adiunkt naukowy
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw, Poland
chrisbuza@yahoo.com

Link prediction with matrix completion techniques

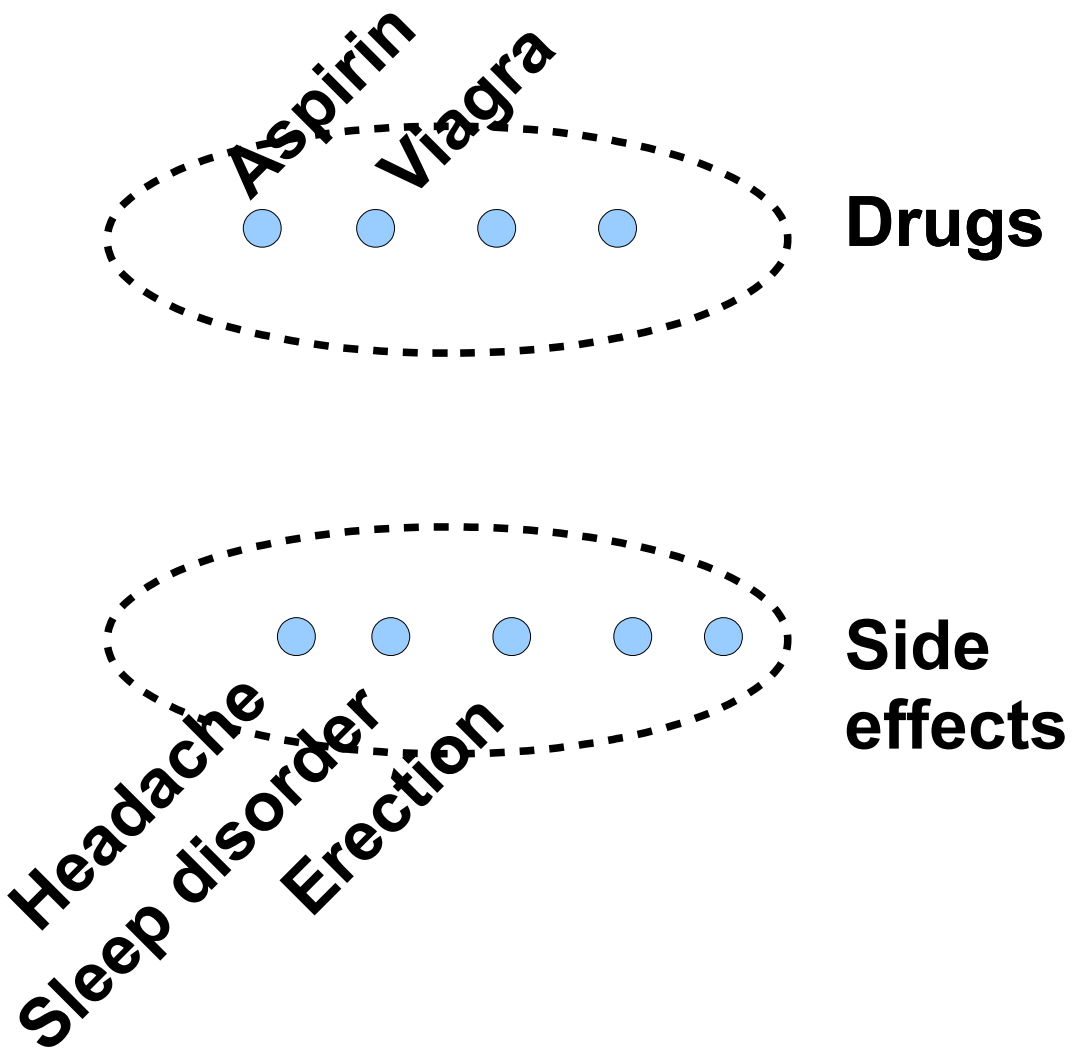
Matrix completion for biomedical tasks

- Drug-target prediction
- Prediction of side effects of drugs
- Link prediction in biological networks
- Analysis of DNA-methylation in case of cancerous tissues

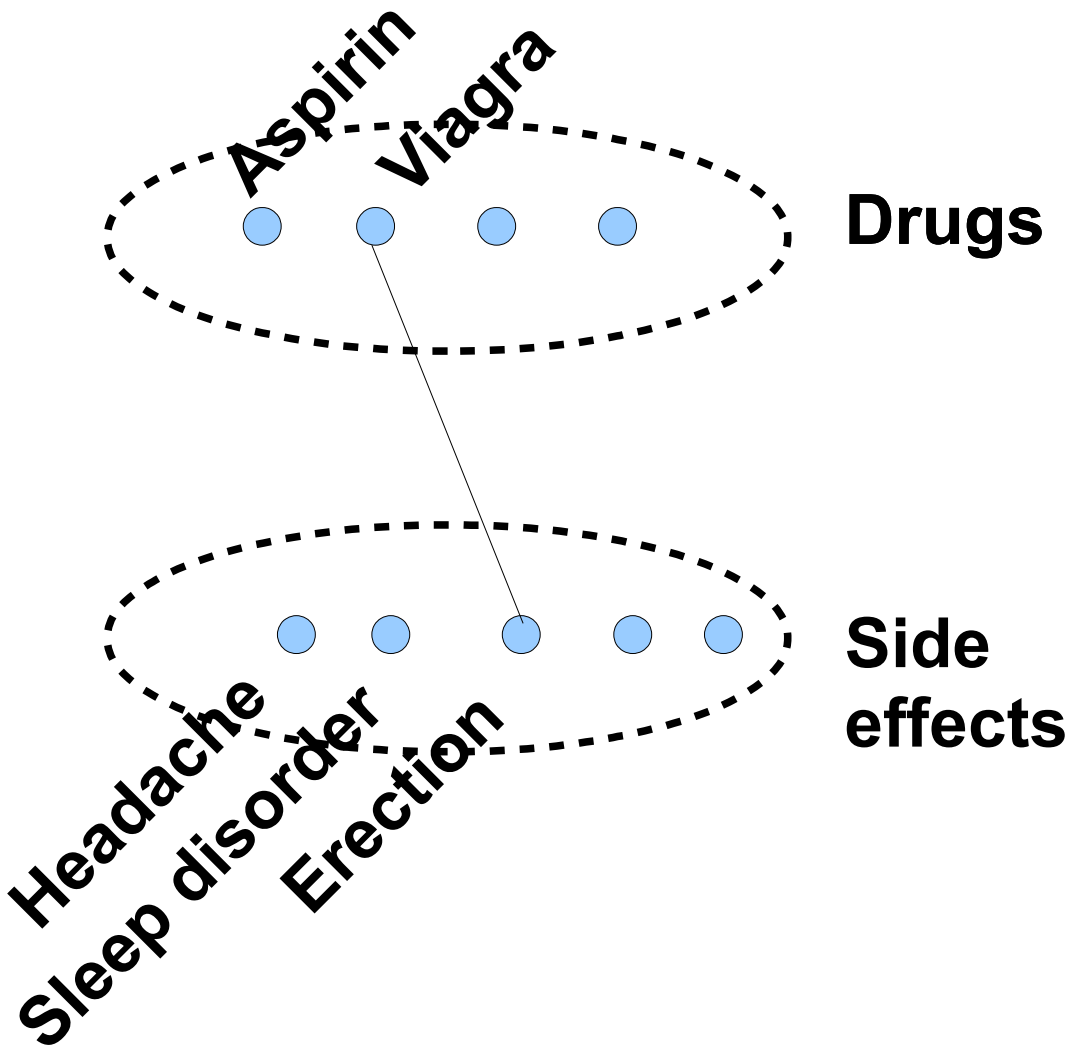
Bipartite graphs



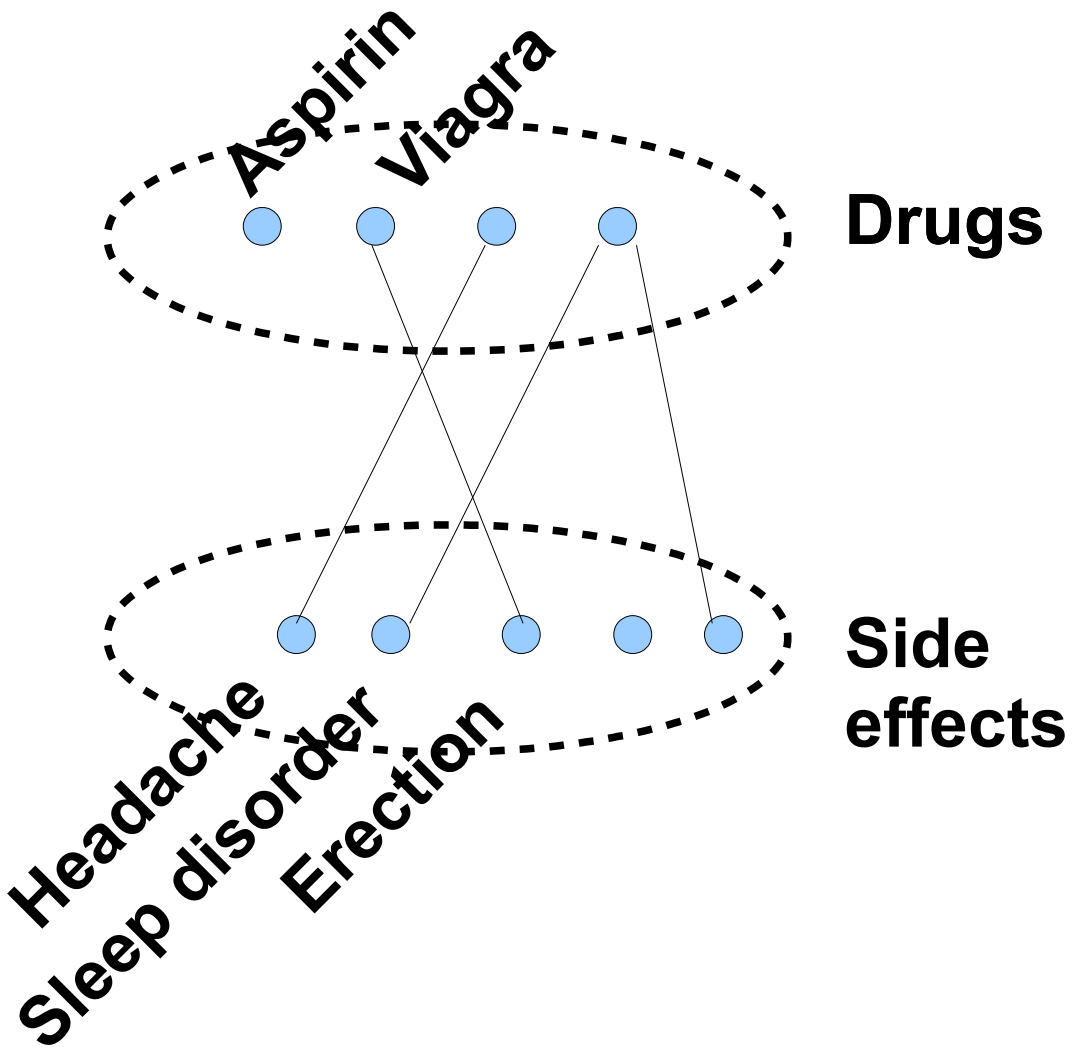
Bipartite graphs



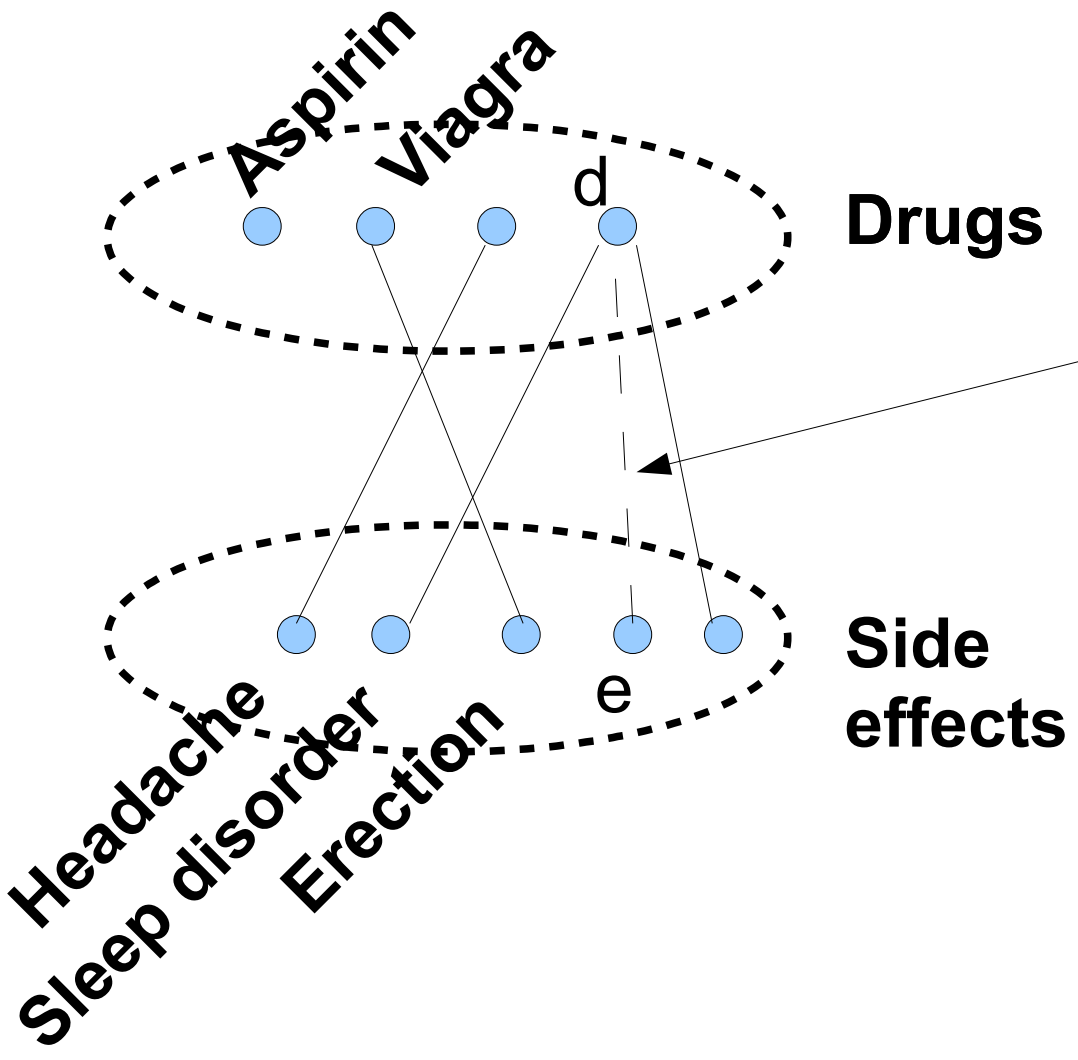
Bipartite graphs



Bipartite graphs



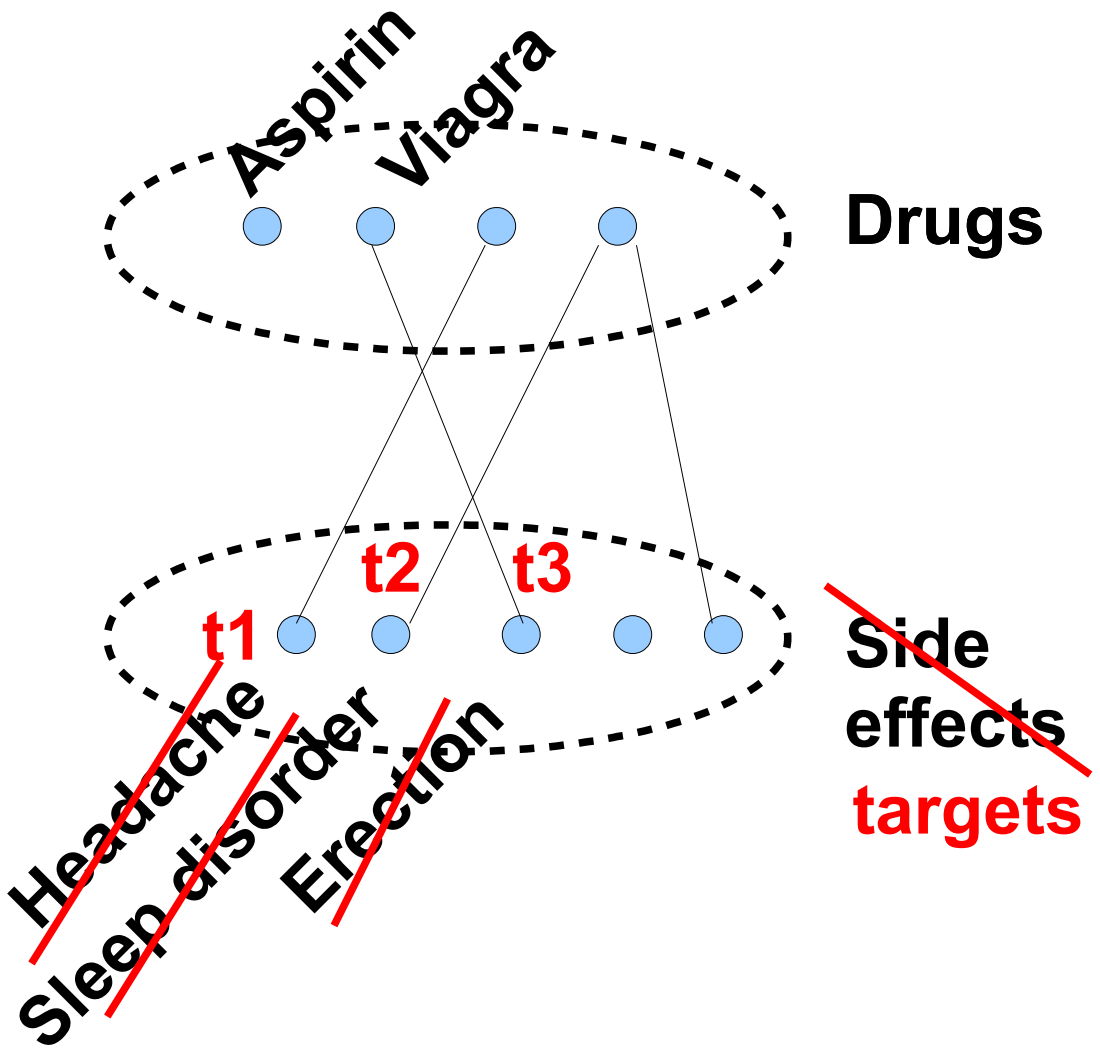
Bipartite graphs



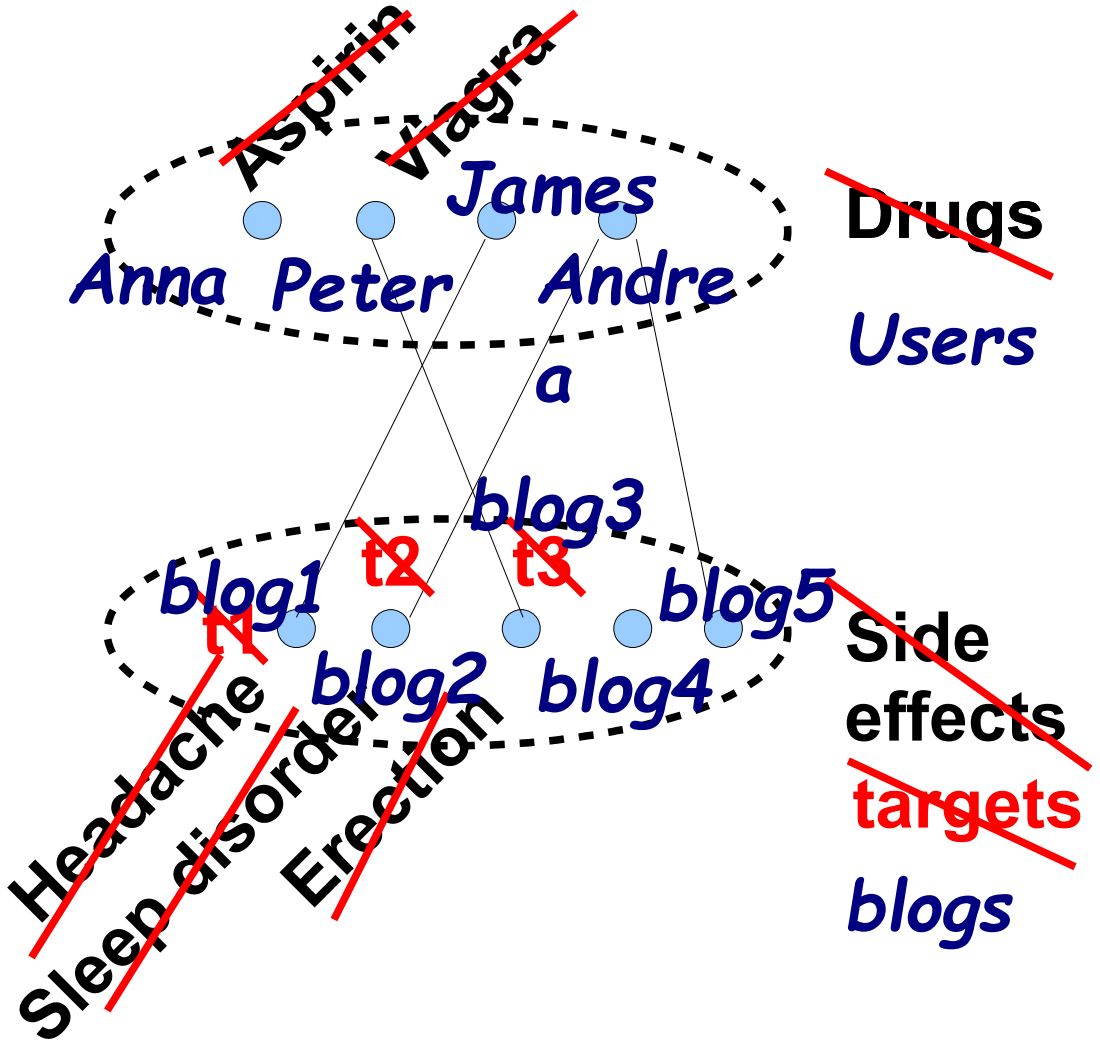
Drug *d* is likely to have side effect *e* (estimated probability: 20%)

Image from Wikipedia is used on this slide. Licence info: http://commons.wikimedia.org/wiki/File:Apple_II.jpg

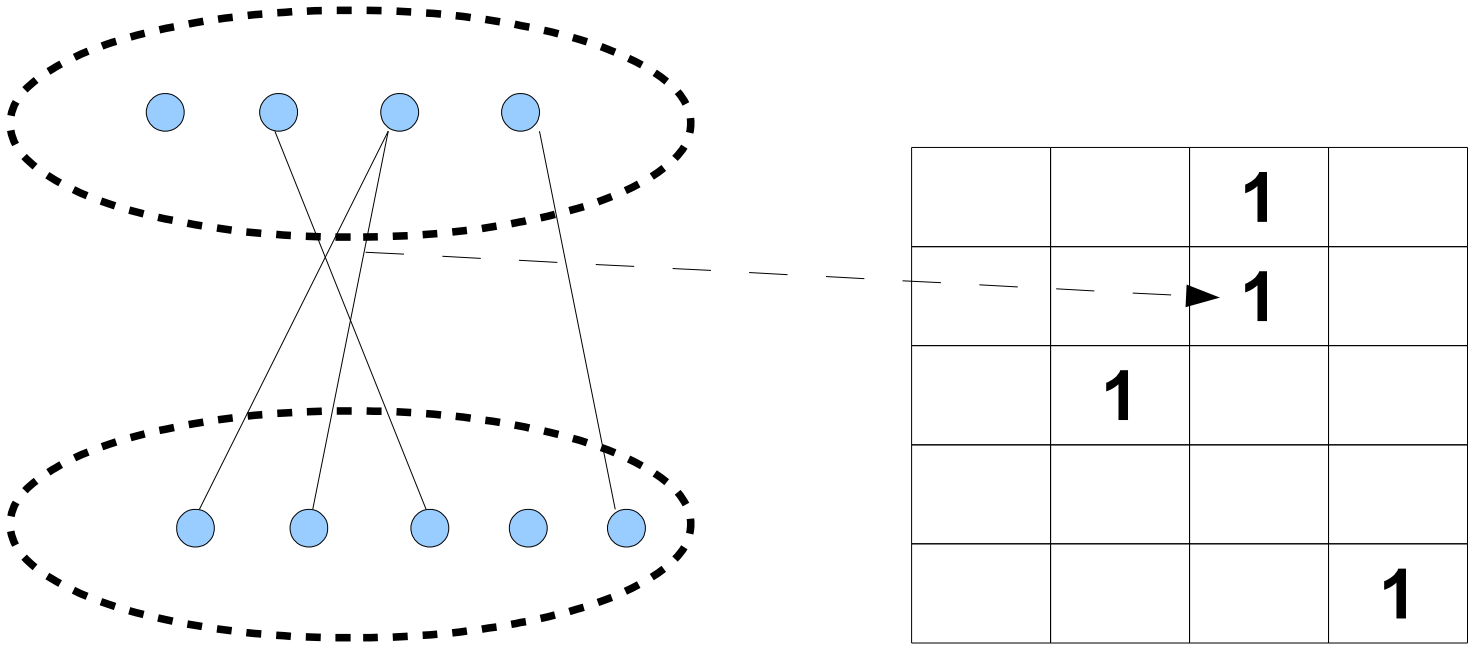
Bipartite graphs



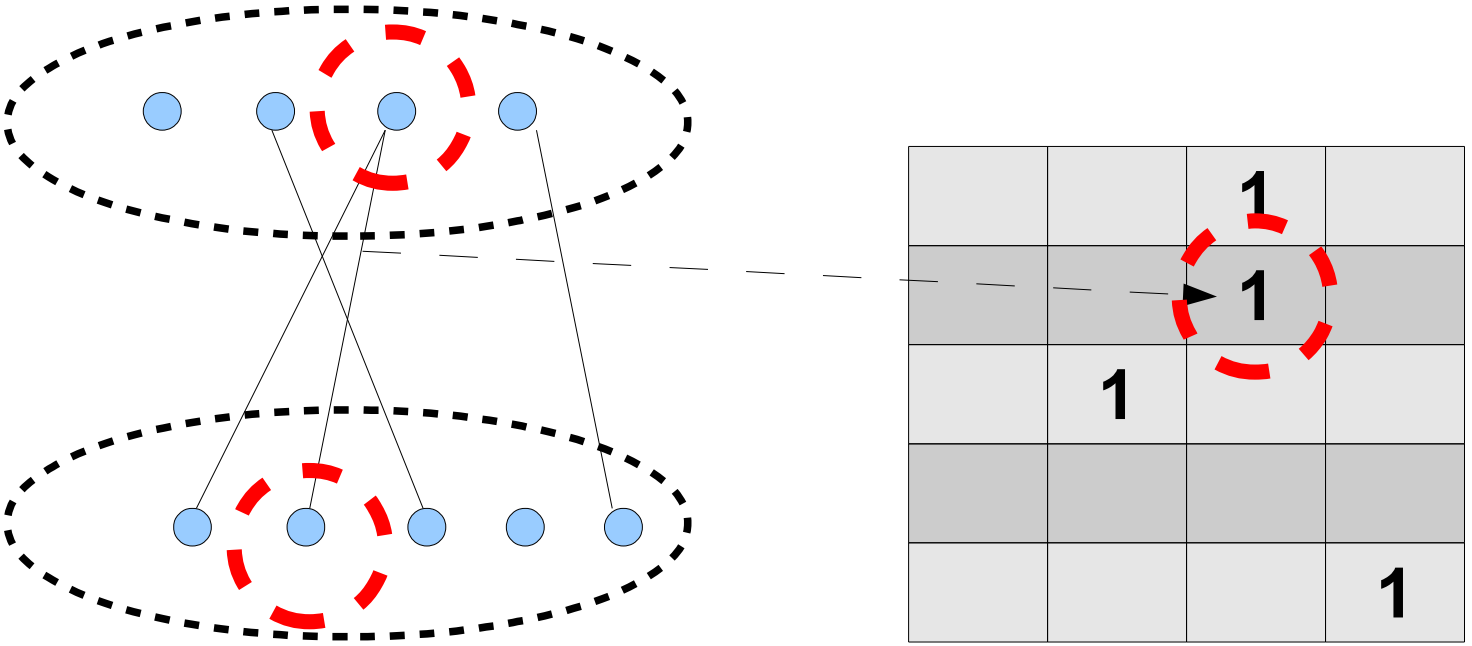
Bipartite graphs



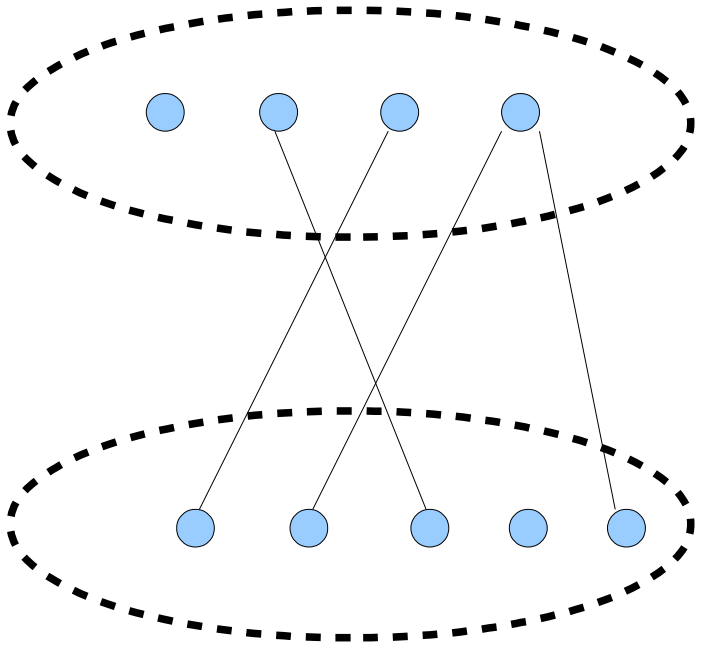
Matrix representation of bipartite graphs

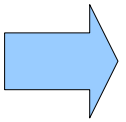


Matrix representation of bipartite graphs



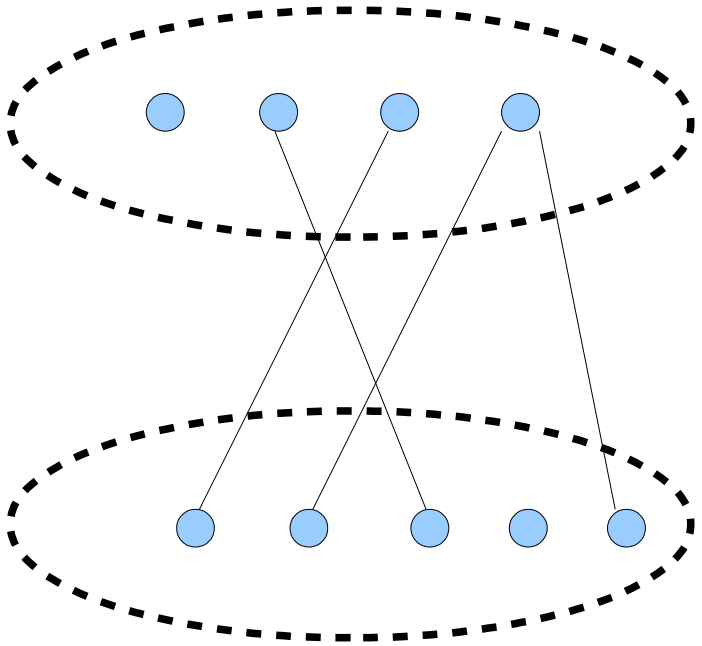
Matrix representation of bipartite graphs

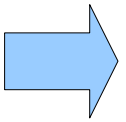




-	-	1	-
-	-	-	1
-	1	-	-
-	-	-	-
-	-	-	1

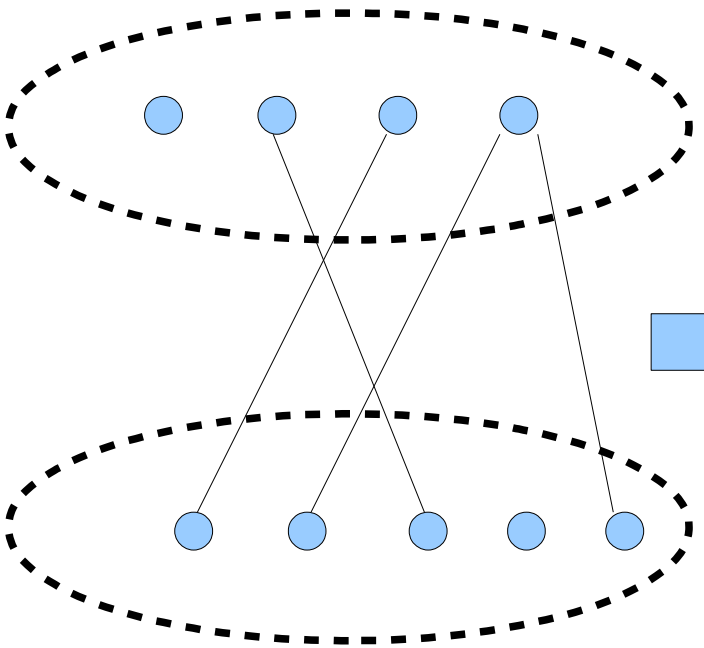
Matrix representation of bipartite graphs





-	-	1	-
0	-	-	1
-	1	0	-
-	-	-	0
-	0	-	1

Matrix factorization for link prediction

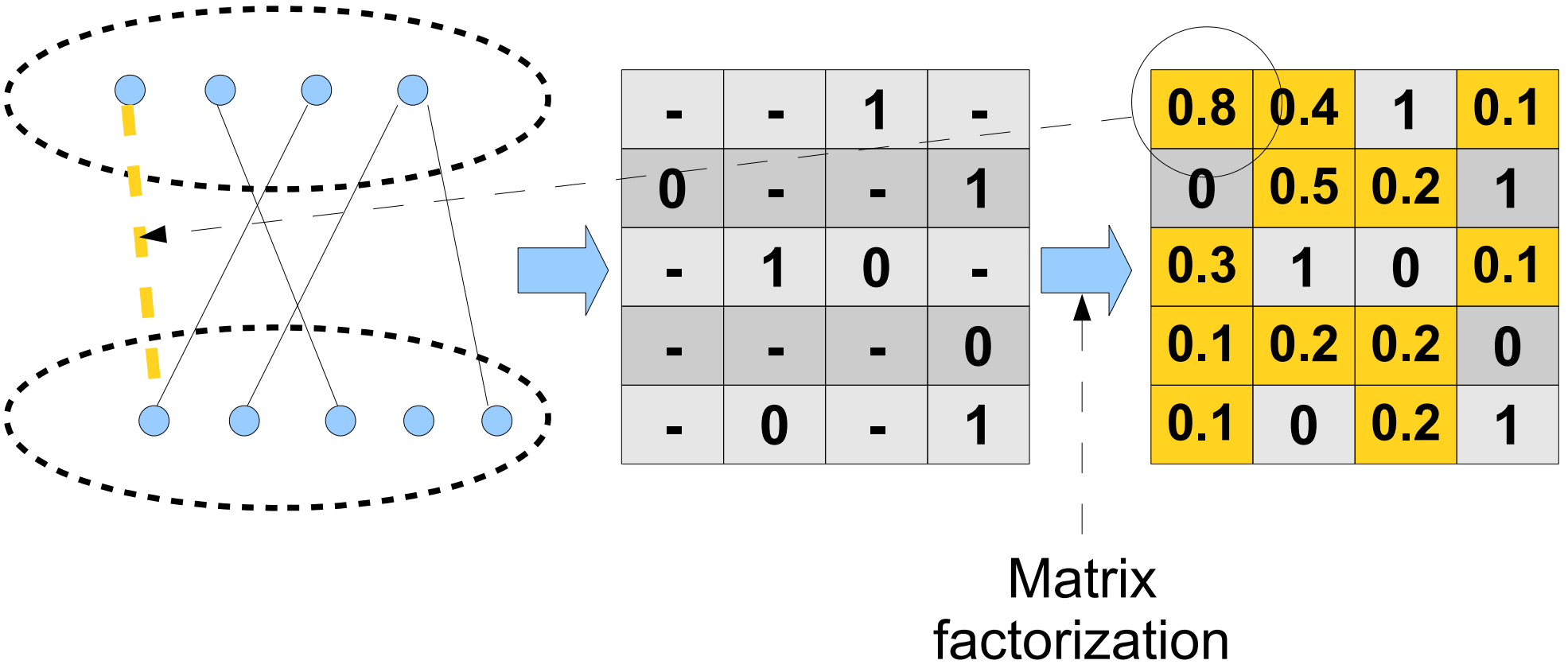


-	-	1	-
0	-	-	1
-	1	0	-
-	-	-	0
-	0	-	1

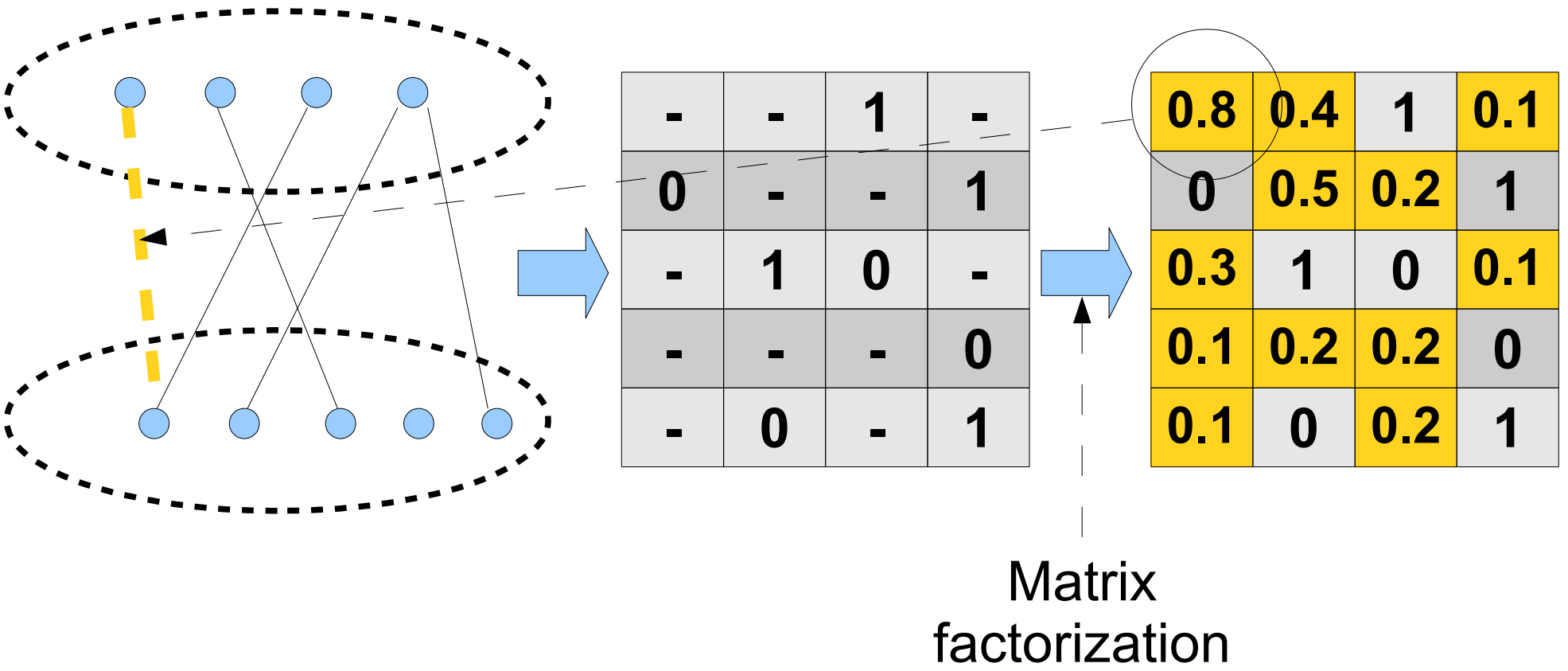
0.8	0.4	1	0.1
0	0.5	0.2	1
0.3	1	0	0.1
0.1	0.2	0.2	0
0.1	0	0.2	1

Matrix factorization

Matrix factorization for link prediction

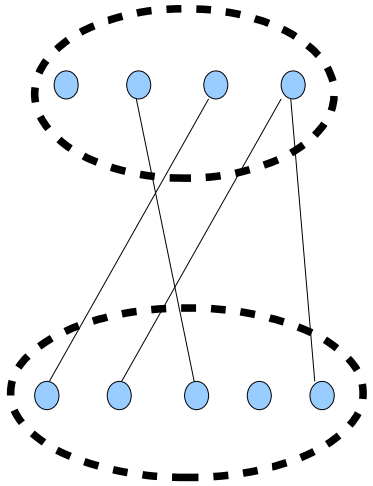


Matrix factorization for link prediction



K. Buza, I. Galambos (2013): An Application of Link Prediction in Bipartite Graphs: Personalized Blog Feedback Prediction, 8th Japanese-Hungarian Symposium on Discrete Mathematics and Its Applications

Combining (slightly) different solutions



-	-	1	-
-	-	-	1
-	1	-	-
-	-	-	-
-	-	-	1

-	-	1	-
0	-	-	1
-	1	0	-
-	-	-	0
-	0	-	1

-	0	1	-
-	-	-	1
0	1	-	-
-	0	-	-
0	-	-	1

-	-	1	0
-	-	0	1
-	1	-	0
0	-	-	-
-	-	-	1

0.8	0.4	1	0.1
0	0.5	0.2	1
0.3	1	0	0.1
0.1	0.1	0.2	0
0.1	0	0.2	1

0.6	0	1	0.3
0.7	0.4	0.2	1
0	1	0.6	0.1
0.1	0	0.3	0.4
0	0.6	0.1	1

0.7	0.6	1	0
0.5	0.6	0	1
0.3	1	0.4	0
0	0.3	0.1	0.2
0.1	0.4	0.3	1

(0.8+0.6+0.7) / 3

0.7	**0.5**	1	**0.2**
0.6	**0.5**	**0.2**	1
0.3	1	**0.5**	**0.1**
0.1	**0.2**	**0.2**	**0.3**
0.1	**0.5**	**0.2**	1

Evaluation of link prediction

- Gold standard = „right solution“
  - Split the data into disjoint train and test sets
  - Split the data into three disjoint sets: train, test1 and test2

Example

- Prediction algorithm predicts 100 news links
- All of them are present in the gold standard
  (all of them are real new links) → performance: 100 %

Example

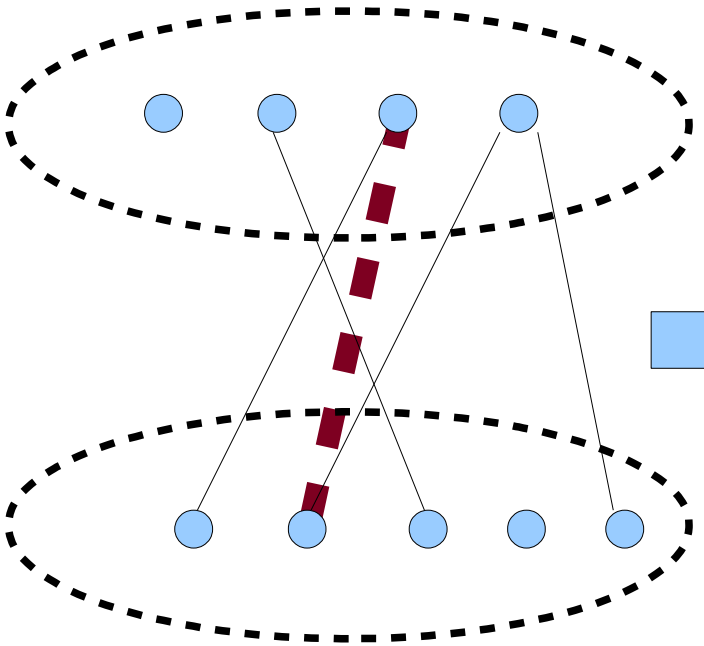
- Prediction algorithm predicts 100 news links
- All of them are present in the gold standard (all of them are real new links) → performance: 100 %
- But: in total, there are 1000 new links in the gold standard (i.e., additionally to the predicted ones, there are 900 other new links in reality) → performance: 10 %

Evaluation of link prediction

- Gold standard = „right solution“
  - Split the data into disjoint train and test sets
  - Split the data into three disjoint sets: train, test1 and test2
- Performance measures
  - Precision:

    P = (# predicted links that are present in gold standard) / (# all predicted links)
  - Recall:

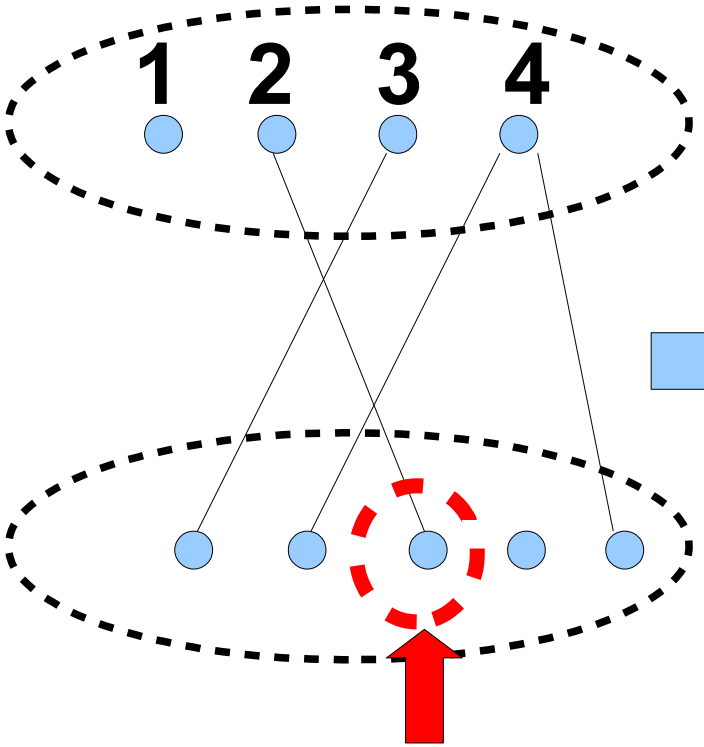
    R = (# predicted links that are present in gold standard) / (# all links in gold standard)
  - F-measure: harmonic mean of precision and recall
    (see also http://en.wikipedia.org/wiki/F1_score)
- Other issues
  - evaluation protocols, statistical significance



-	-	1	-
0	-	-	1
-	1	0	-
-	-	-	0
-	0	-	1

0.8	0.4	1	0.1
0	0.5	0.2	1
0.3	1	0	0.1
0.1	0.2	0.2	0
0.1	0	0.2	1

Matrix factorization



-	-	1	-
0	-	-	1
-	1	0	-
-	-	-	0
-	0	-	1

0.8	0.4	1	0.1
0	0.5	0.2	1
0.3	1	0	0.1
0.1	0.2	0.2	0
0.1	0	0.2	1

Matrix factorization