Unsupervised Entity Resolution on Multi-type Graphs

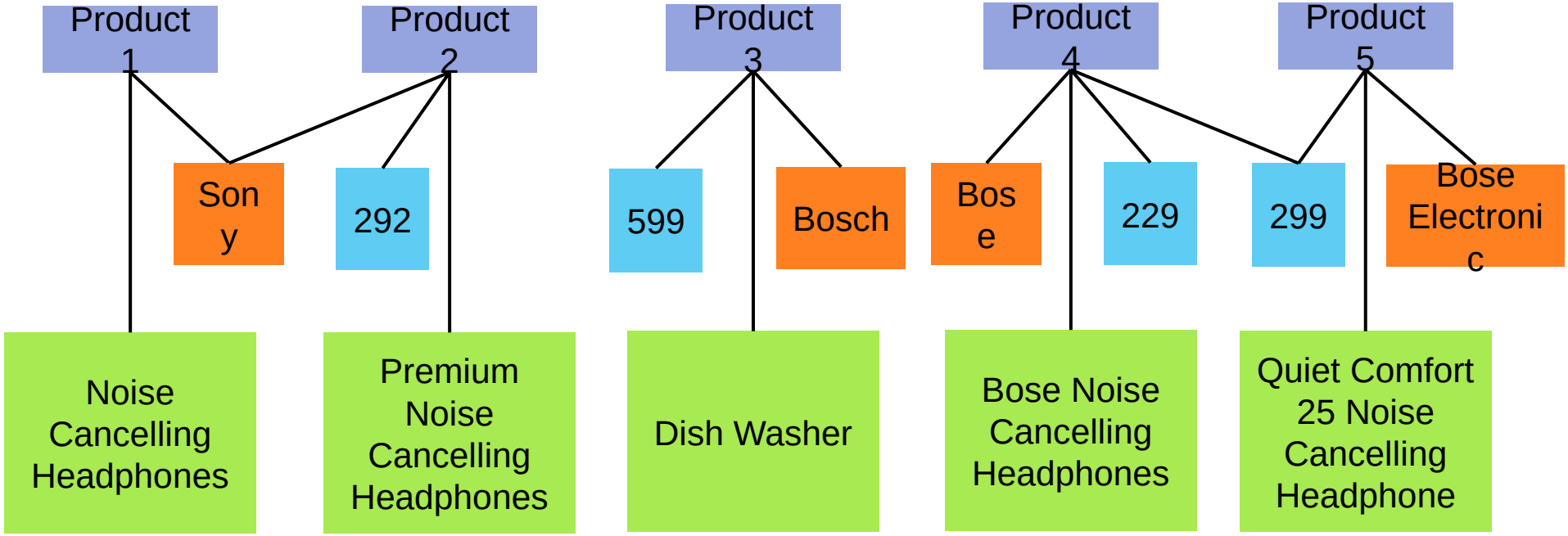
Linhong Zhu, Majid Ghasemi-Gol,

Pedro Szekely, Aram Galstyan, Craig A. Knoblock

Information Sciences Institute, University of Southern California

Entity Resolution

Identifying and linking instances of the same real world entity



product

manufacturer

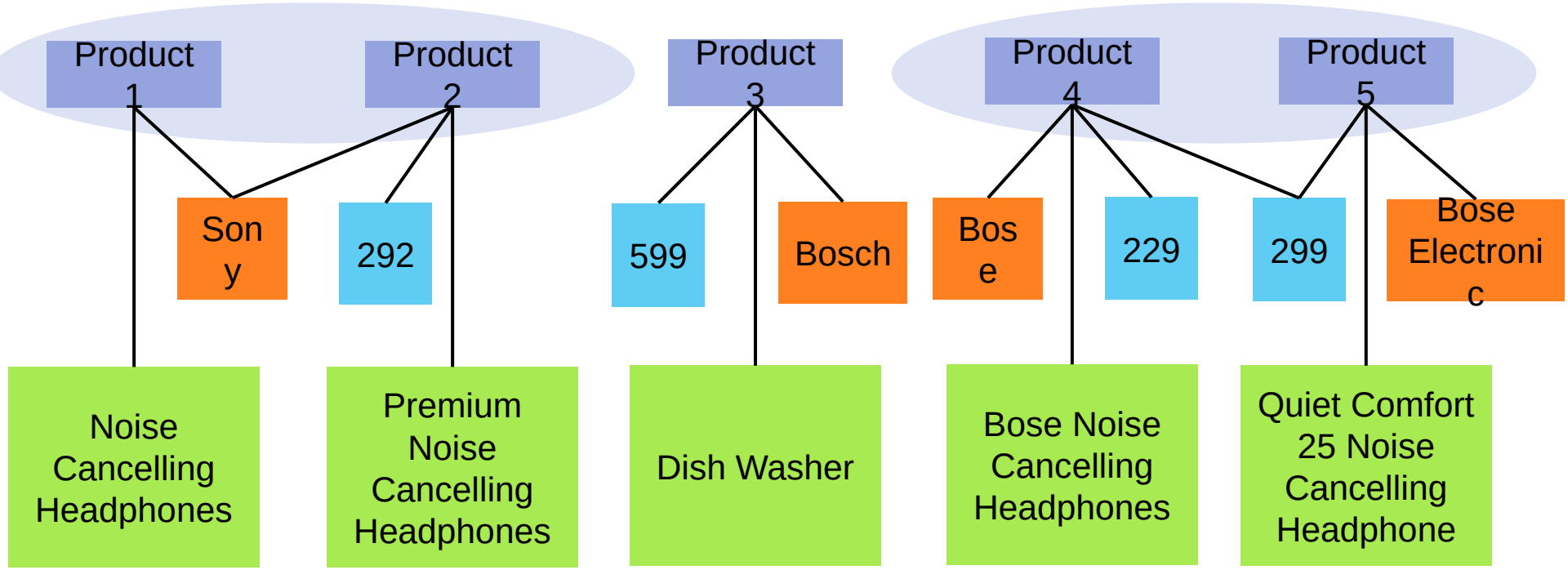
price

description

Multi-Type Graph

Entity Resolution

Identifying and linking instances of the same real world entity



product

manufacturer

price

description

Multi-Type Graph

Common Approach: Pairwise Comparisons

	Manufacturer	Price	Title
Product 1	Sony		Noise Cancelling Headphones
Product 2	Sony	292	Premium Noise Cancelling Headphones
Product 3	Bosch	599	Dish Washer
Product 4	Bose	299, 229	Bose Noise Cancelling Headphones
Product 5	Bose Electronic	299	Quiet Comfort 25 Noise Cancelling Headphone

Acceptance Threshold: 0.8

Missing Values

	Manufacturer	Price	Title
Product 1	Sony		Noise Cancelling Headphones
Product 2	Sony	292	Premium Noise Cancelling Headphones
Product 3	Bosch	599	Dish Washer
Product 4	Bose	299, 229	Bose Noise Cancelling Headphones
Product 5	Bose Electronic	299	Quiet Comfort 25 Noise Cancelling Headphone

Multiple Values

	Manufacturer	Price	Title
Product 1	Sony		Noise Cancelling Headphones
Product 2	Sony	292	Premium Noise Cancelling Headphones
Product 3	Bosch	599	Dish Washer
Product 4	Bose	299, 229	Bose Noise Cancelling Headphones
Product 5	Bose Electronic	299	Quiet Comfort 25 Noise Cancelling Headphone

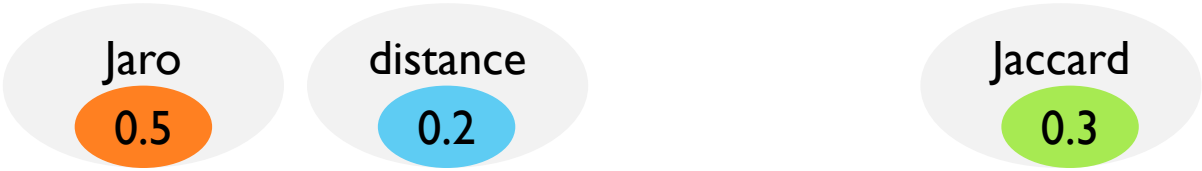
Jaro
0.5

distance
0.2

Jaccard
0.3

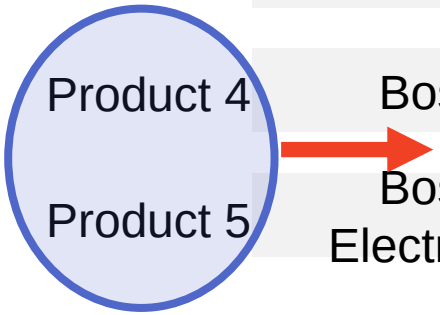
Weights

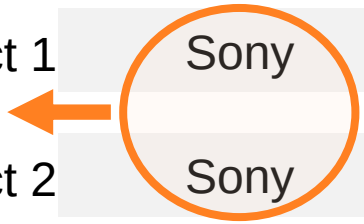
	Manufacturer	Price	Title
Product 1	Sony		Noise Cancelling Headphones
Product 2	Sony	292	Premium Noise Cancelling Headphones
Product 3	Bosch	599	Dish Washer
Product 4	Bose	299, 229	Bose Noise Cancelling Headphones
Product 5	Bose Electronic	299	Quiet Comfort 25 Noise Cancelling Headphone



Unidirectional

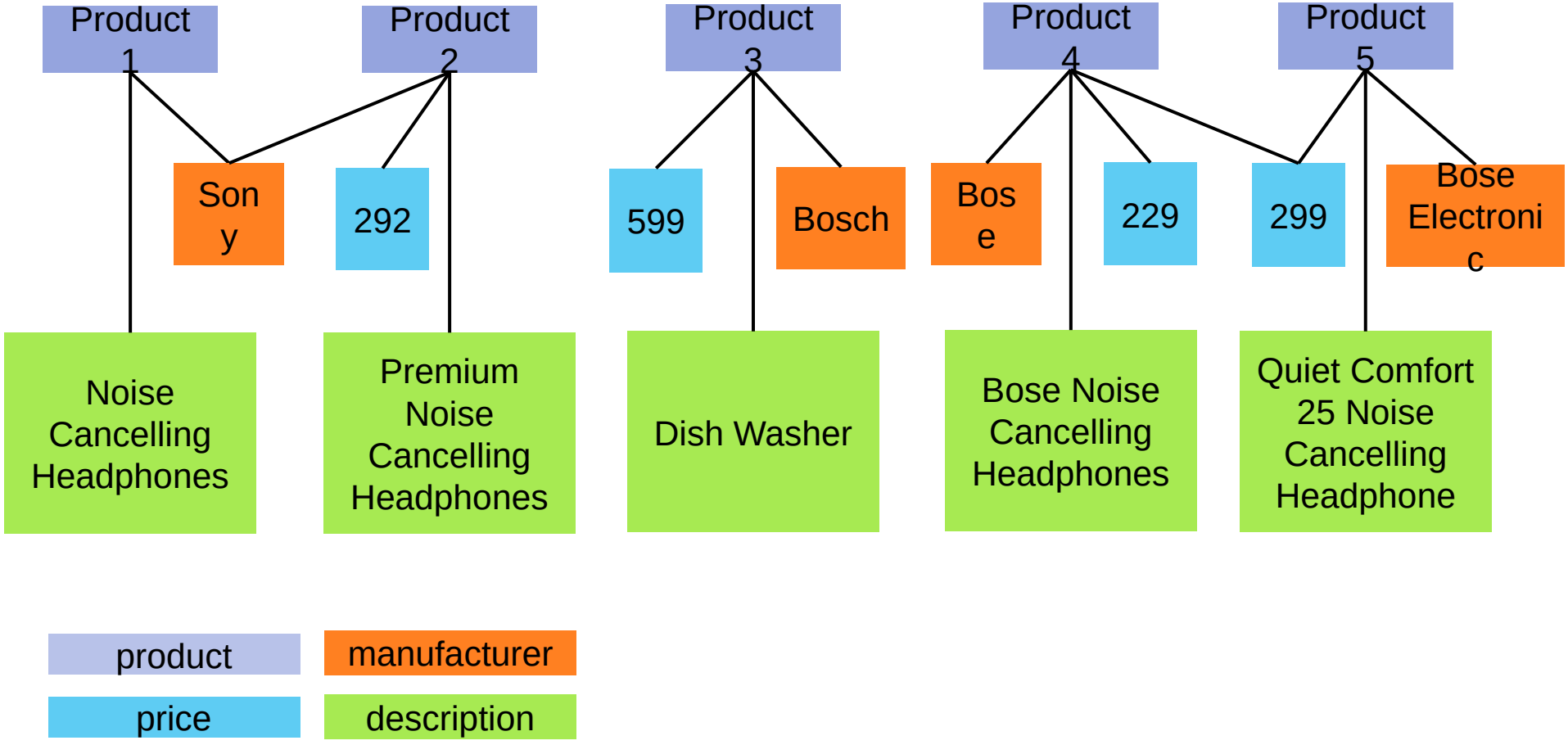
	Manufacturer	Price	Title
Product 1	Sony		Noise Cancelling Headphones
Product 2	Sony	292	Premium Noise Cancelling Headphones
Product 3	Bosch	599	Dish Washer
Product 4	Bose	299, 229	Bose Noise Cancelling Headphones
Product 5	Bose Electronic	299	Quiet Comfort 25 Noise Cancelling Headphone





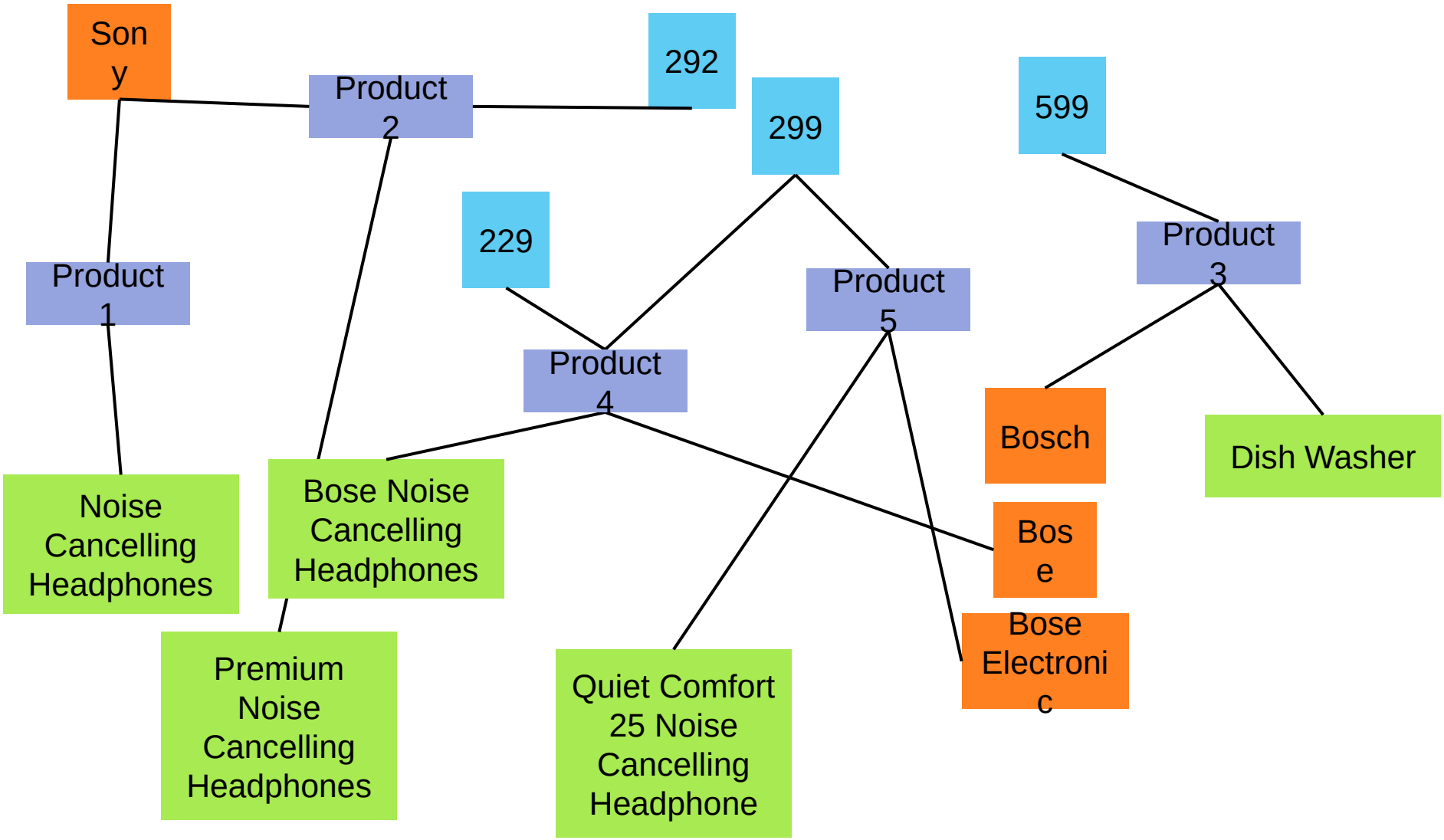
Jaro	distance	Jaccard
0.5	0.2	0.3

Graph Summarization:
Original Graph

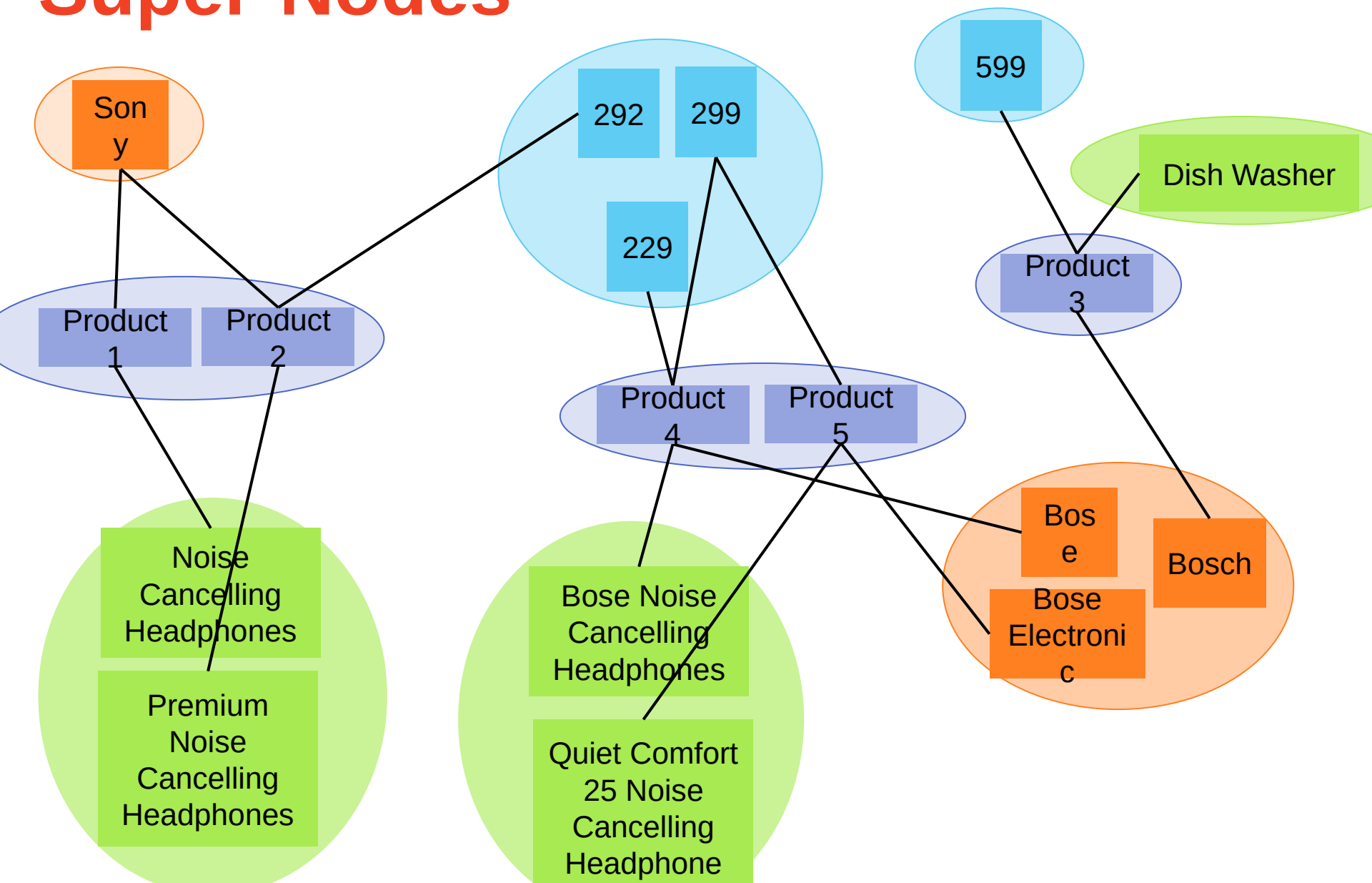


Super-Nodes

Similar Nodes $sim_t(x, y)$



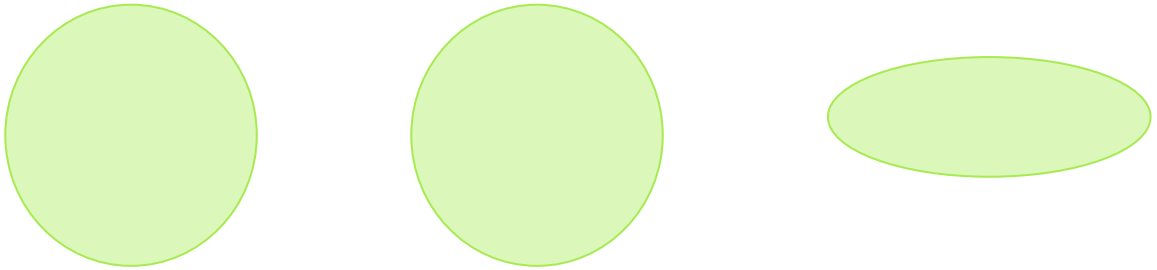
Graph Sumarization:
Super-Nodes



Super-nodes $C_t(x)$

probability that a node x belongs to each super-node
one matrix for each type

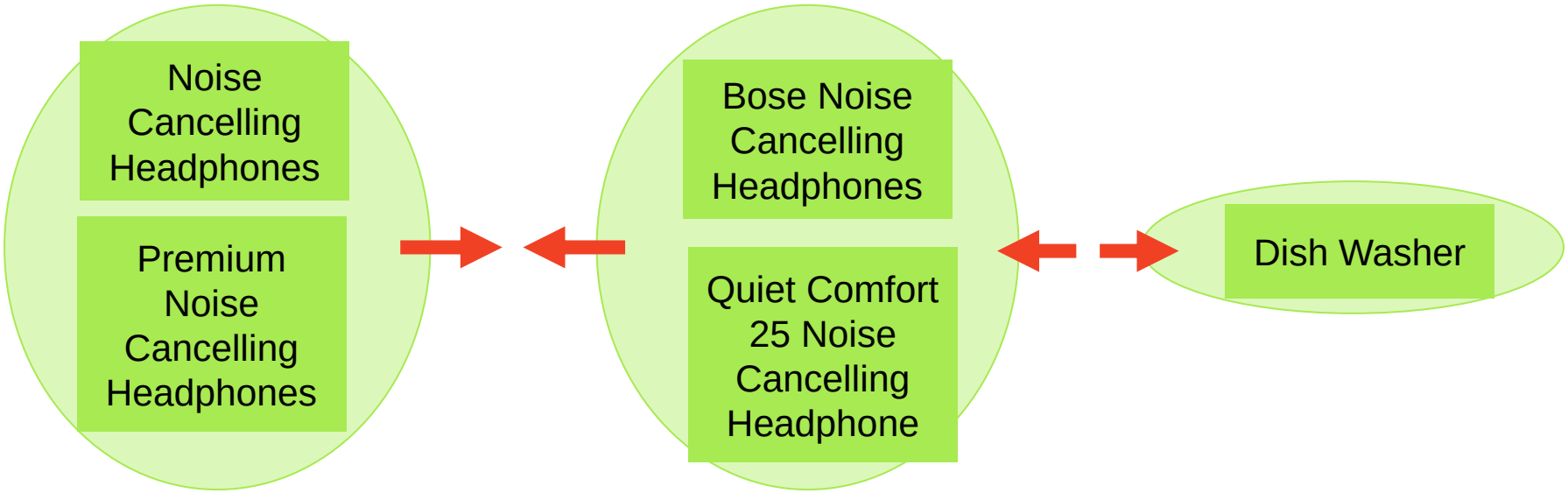
C_t



Noise Cancelling Headphones	**0.7**	0.2	0.1
Premium Noise Cancelling Headphones	**0.7**	0.2	0.1
Bose Noise Cancelling Headphones	0.2	**0.7**	0.1
Quiet Comfort 25 Noise Cancelling Headphone	0.2	**0.7**	0.1
Dish Washer	0.1	0.1	**0.8**

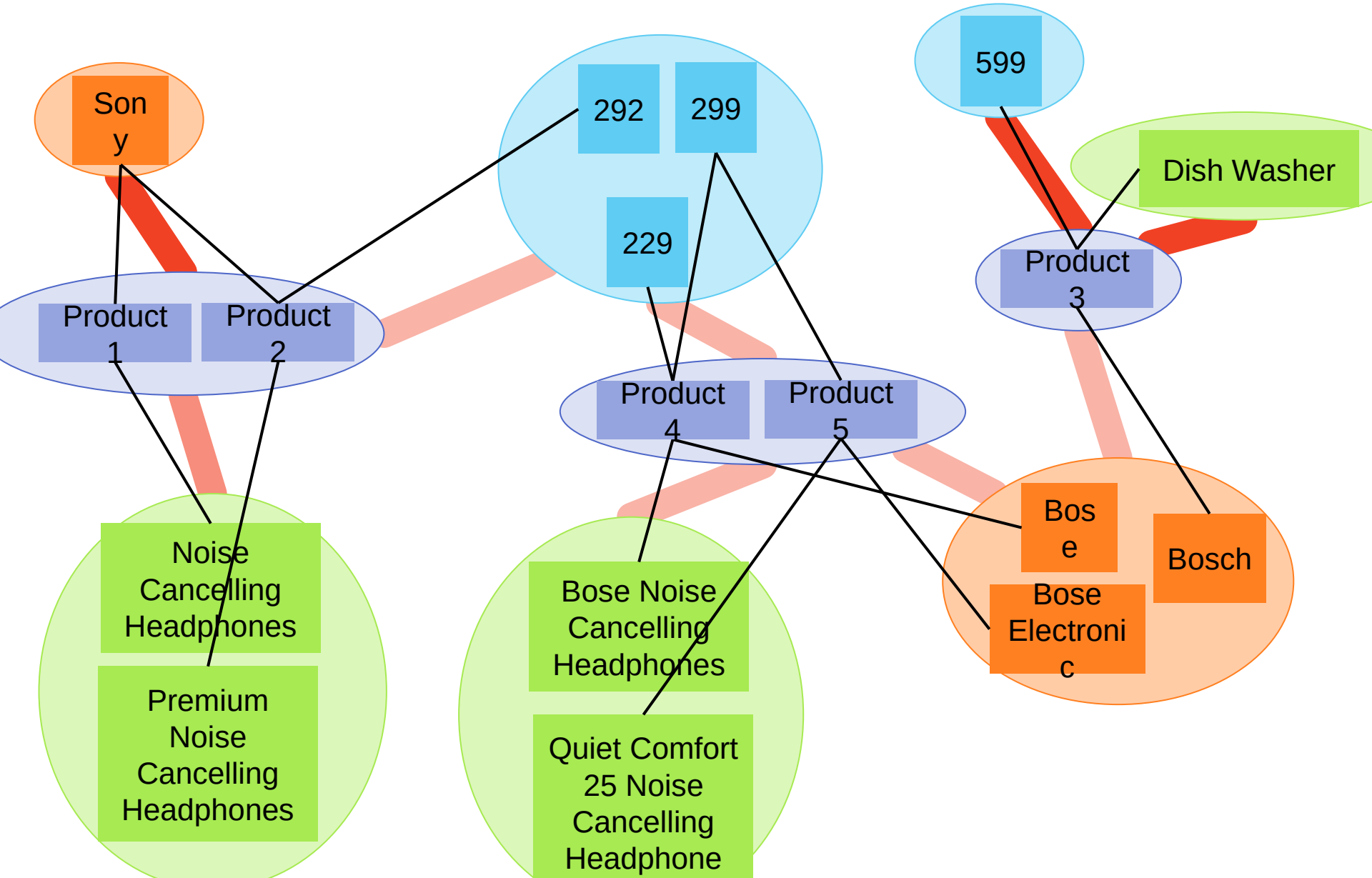
Similar Nodes Should Be In The Same Super-Node

$$\sum_{t} \sum_{x,y \in V_t} \mathrm{sim}_t(x,y) \|C_t(x) - C_t(y)\|_F^2$$



Super-Links

Super-Links


Son
y
292
299
229
599
Dish Washer
Product
1
Product
2
Product
3
Product
4
Product
5
Noise
Cancelling
Headphones
Premium
Noise
Cancelling
Headphones
Bose Noise
Cancelling
Headphones
Quiet Comfort
25 Noise
Cancelling
Headphone
Bos
e
Bosch
Bose
Electroni
c

Reconstruct Original Graph From Summary Graph

Original Graph

Summary Graph

$$\sum_{t} \sum_{t' > t} \| G_{tt'} - C_t L_{tt'} {C_{t'}}^T \|_F^2$$

Graph Types

Graph Summarization

Original Graph

Summary Graph

$$\sum_{t}\sum_{t'>t} \| G_{tt'} - C_t L_{tt'} C_{t'}^{T} \|_F^2$$

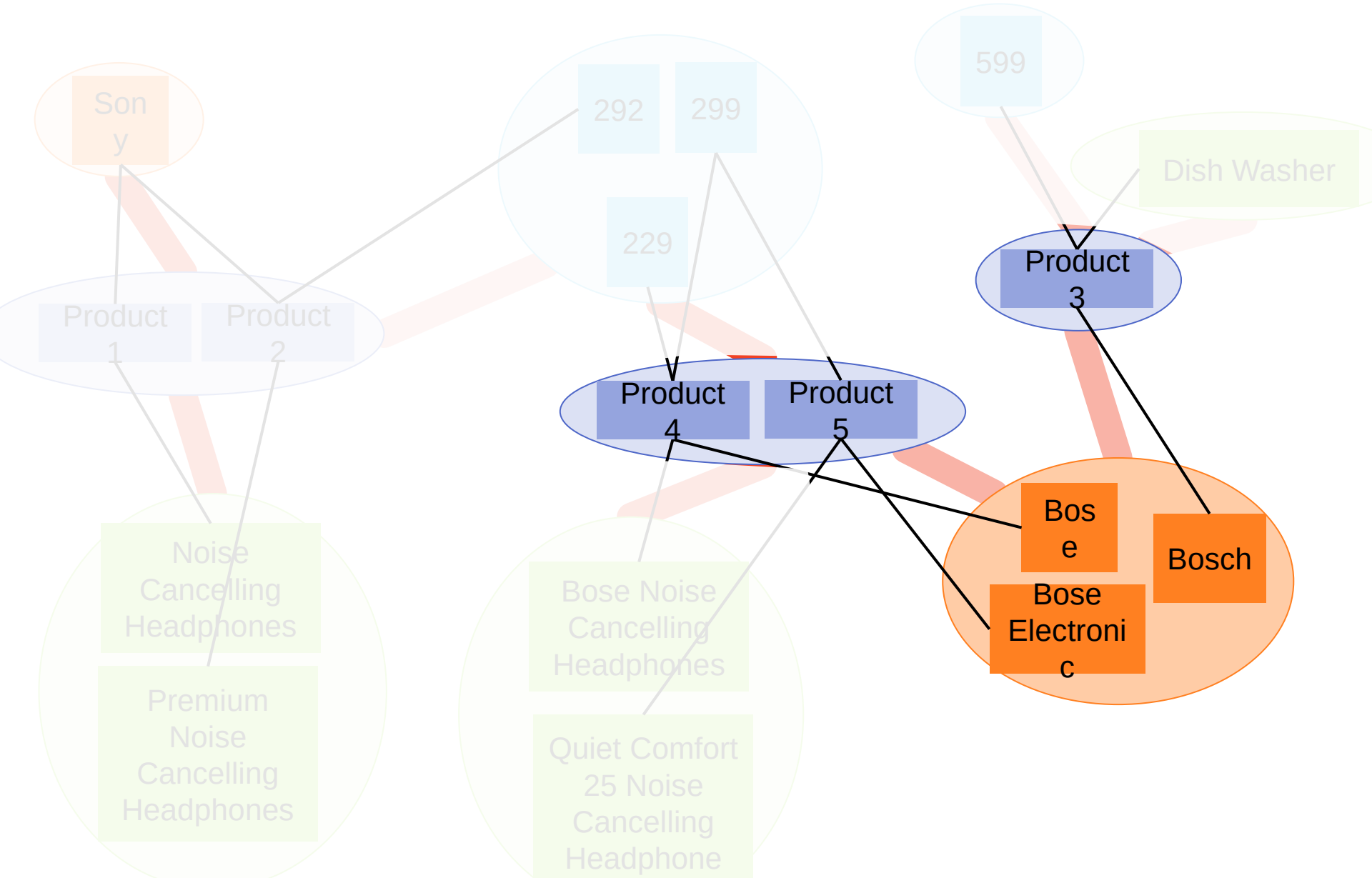
Graph Types

Super Links

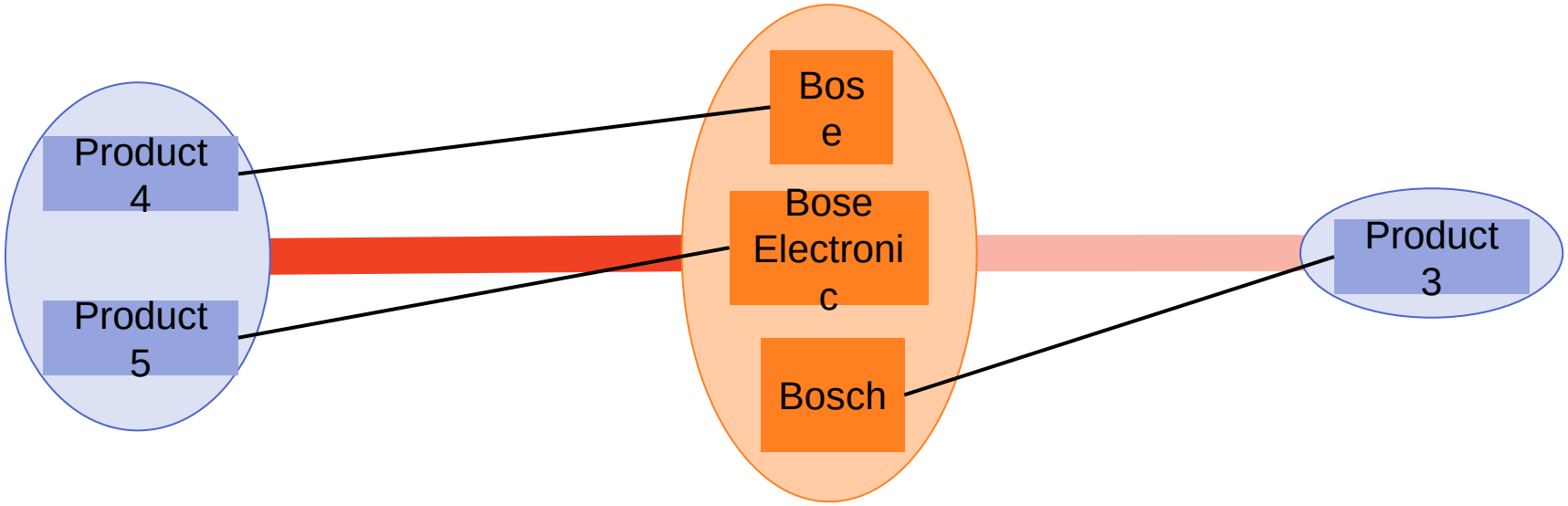
Super-nodes of t-nodes

Super-nodes of t'-nodes

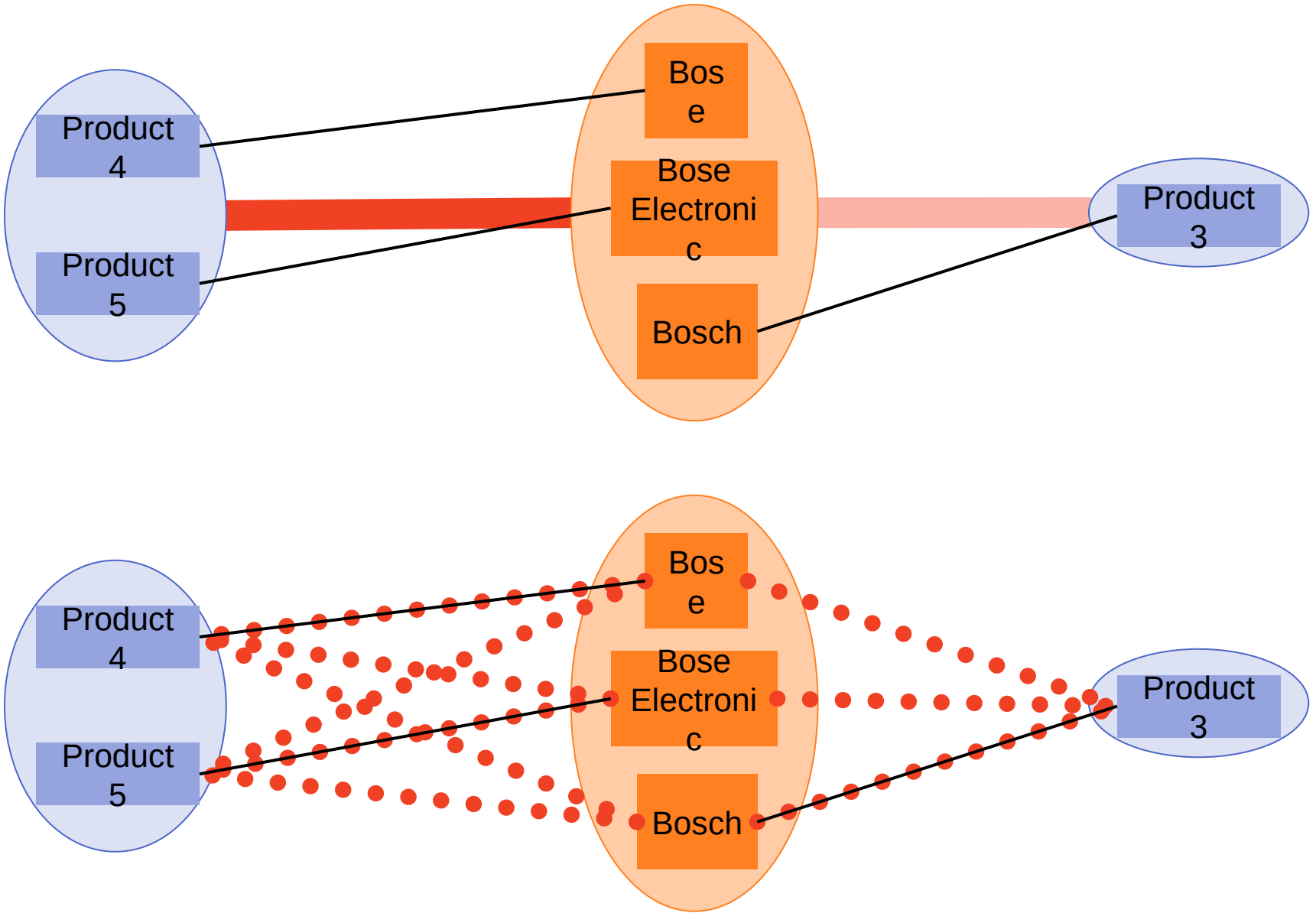
Super-Links


Son
y
292
299
599
Dish Washer
229
Product
3
Product
1
Product
2
Product
4
Product
5
Bos
e
Bosch
Bose
Electroni
c
Noise
Cancelling
Headphones
Premium
Noise
Cancelling
Headphones
Bose Noise
Cancelling
Headphones
Quiet Comfort
25 Noise
Cancelling
Headphone

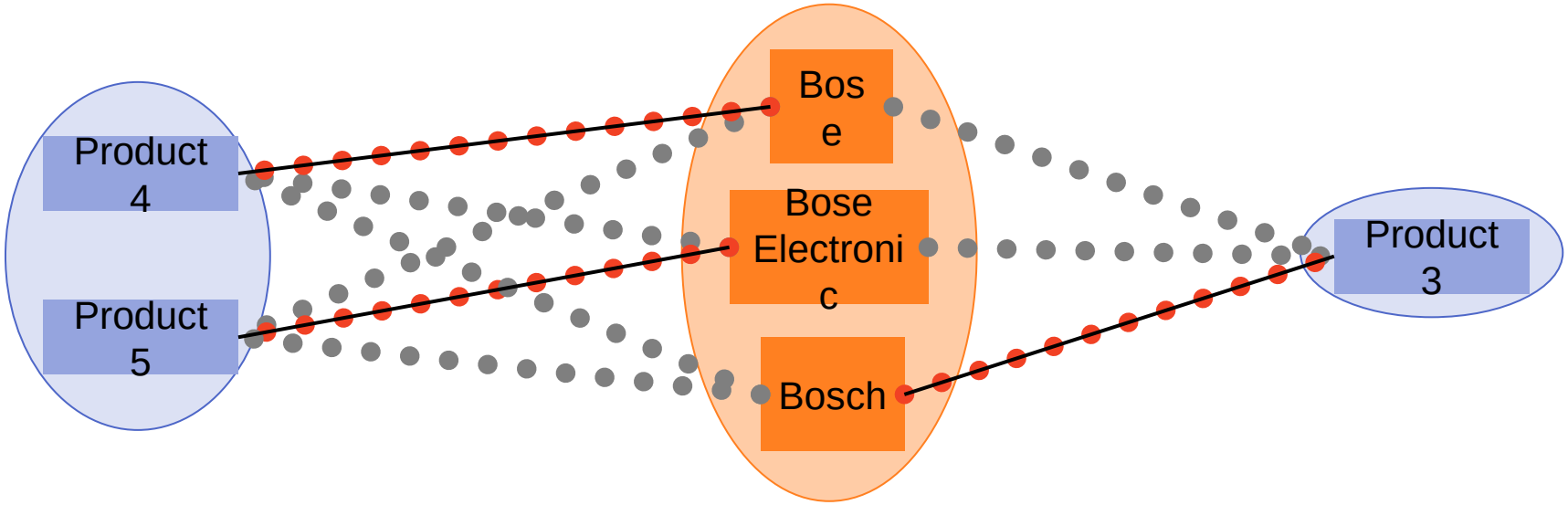
Predict Links In Original Graph



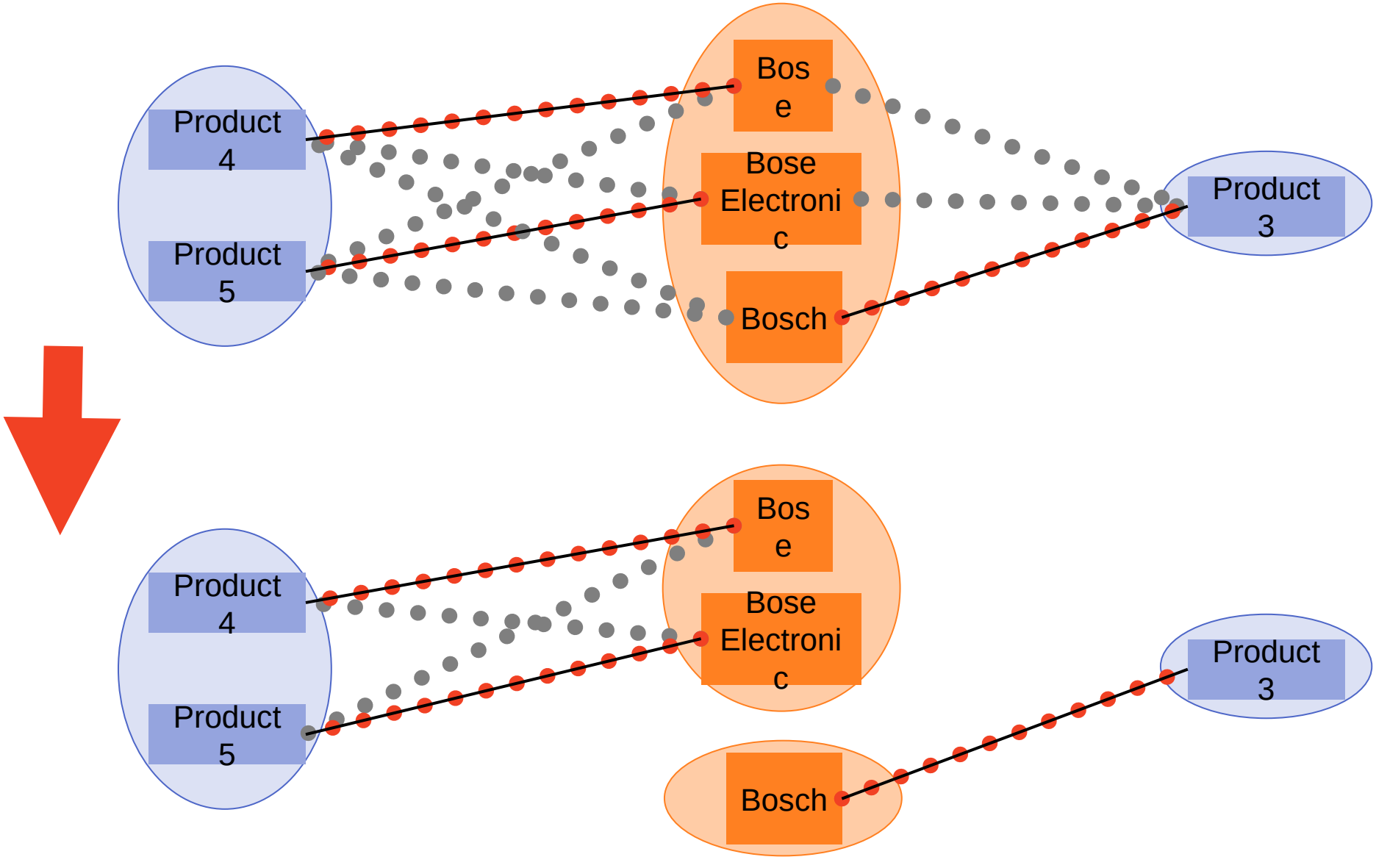
Predict Links In Original Graph



Predict Links In Original Graph



Re-Clustering Improves Reconstruction Quality



Minimization Problem

super-node coherence

$$\sum_{t} \sum_{x,y \in V_t} \mathtt{sim}_t(x,y) \| C_t(x) - C_t(y) \|_F^2$$

$$+ \sum_{t} \sum_{t' > t} \| G_{tt'} - C_t L_{tt'} C_{t'}{}^T \|_F^2$$

structural coherence

Iterative Algorithm

re-cluster

$$C_t = C_t \circ \sqrt{\frac{\sum_{t'>t} G_{tt'} C_{t'} {L_{tt'}}^T + \mathtt{sim}_t C_t}{\sum_{t'>t} C_t E_{tt'} {C_{t'}}^T C_{t'} {E_{tt'}}^T + D_t C_t}}$$

$$L_{tt'} = L_{tt'} \circ \sqrt{\frac{{C_t}^T G_{tt'} C_{t'}}{C_t^T C_t L_{tt'} {C_{t'}}^T C_{t'}}}$$

recompute super-links

Evaluation

DataSets

Citeseer

Cluster authors

Cluster publications

Product

Match products from Amazon and Google

Data	# types	# records	# nodes	# edges	# entities	Full input mapping
Citeseer	4	2,892	8,591	17,521	1,165 authors 899 papers	8.4 Million
Product	2	4,589	12,397	41,165	1,104 products	4.4 Million

Comparable Approaches

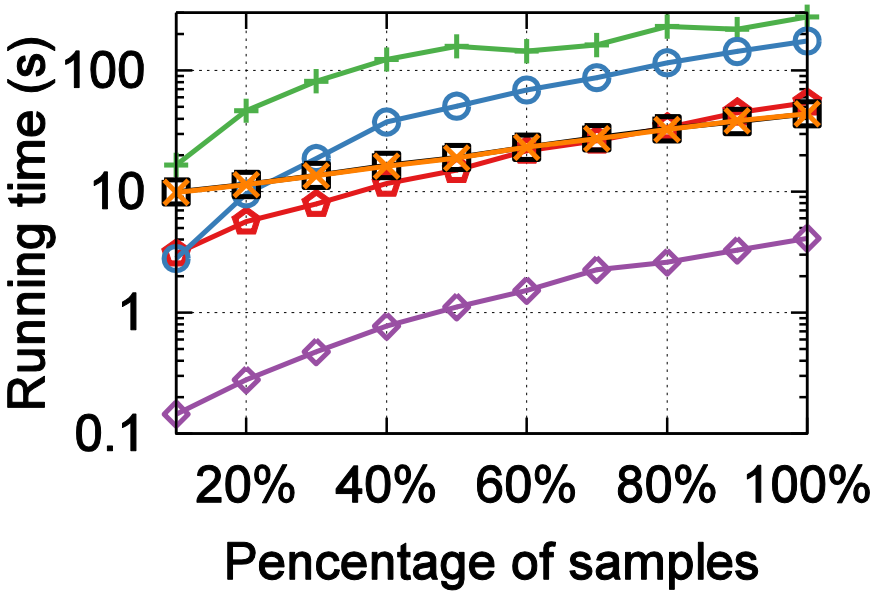
	Pairwise	Clustering	Unsupervised	Supervised
Limes, Ngomo'11	✔		✔	
SILK, Isele'10	✔		✔	✔
Serf, Benjelloun'10	✔		✔	
***Commercial**, Köpcke'10	✔			✔
GraphSum, Riondato'14		✔	✔	
***AuthorLDA**, Bhattacharya'07		✔	✔	
CoSum (proposed)		✔	✔	

Quality Comparison

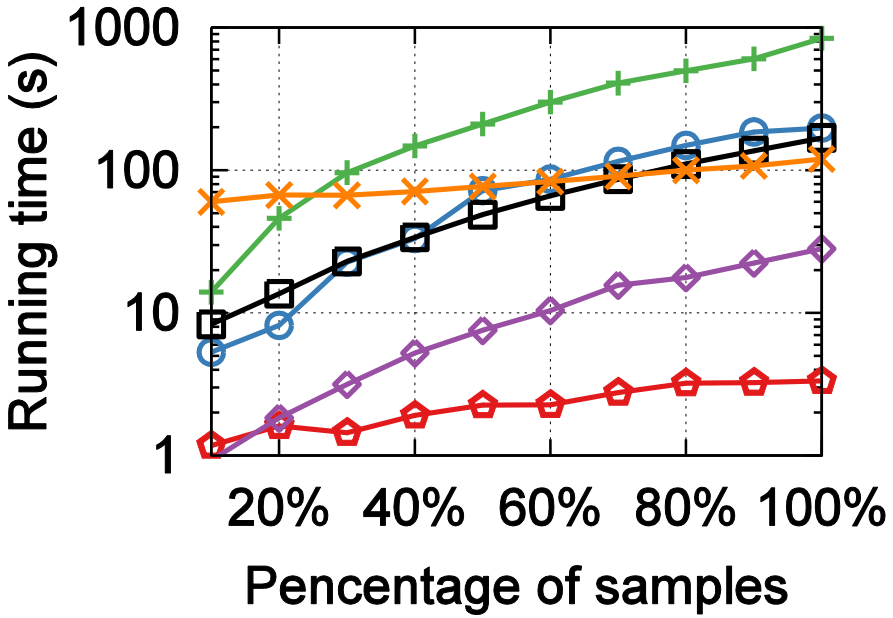
	Precision			Recall			F-measure		
	Author	**Paper**	**Product**	**Author**	**Paper**	**Product**	**Author**	**Paper**	**Product**
Limes-F	0.958	0.827	0.446	0.864	0.761	0.16	0.909	0.792	0.236
Silk-F	0.846	**0.877**	0.459	0.986	0.756	0.348	0.91	0.812	0.395
Gsum	0.727	0.668	0.01	0.569	0.624	0.587	0.638	0.645	0.02
CoSum-B	0.993	0.871	0.58	0.94	0.611	0.477	0.966	0.718	0.524
Limes-MO	0.912	0.827	0.446	0.944	0.761	0.16	0.928	0.792	0.236
Silk-MO	0.932	**0.877**	0.459	0.958	0.756	0.348	0.945	0.812	0.395
Serf	0.985	0.837	0.436	0.687	0.808	0.186	0.809	0.822	0.261
CoSum-P	**0.999**	0.771	**0.639**	**0.997**	**0.997**	**0.695**	**0.998**	**0.87**	**0.666**
Commercial			0.615			0.63			0.622
AuthorLDA							0.995		

Efficiency Comparison

Citeseer



Product



Limes Silk Serf GSum CoSum-B CoSum-P

Conclusion

Graph summarization for entity resolution

Supports many-to-many relations and missing values

No tuning of weights or thresholds

Multi-directional flow of information in the graph

Good results on two benchmark data sets

Future work

Enhance scalability for dense graphs

System engineering

Semi-supervised approach to speed up convergence