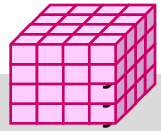
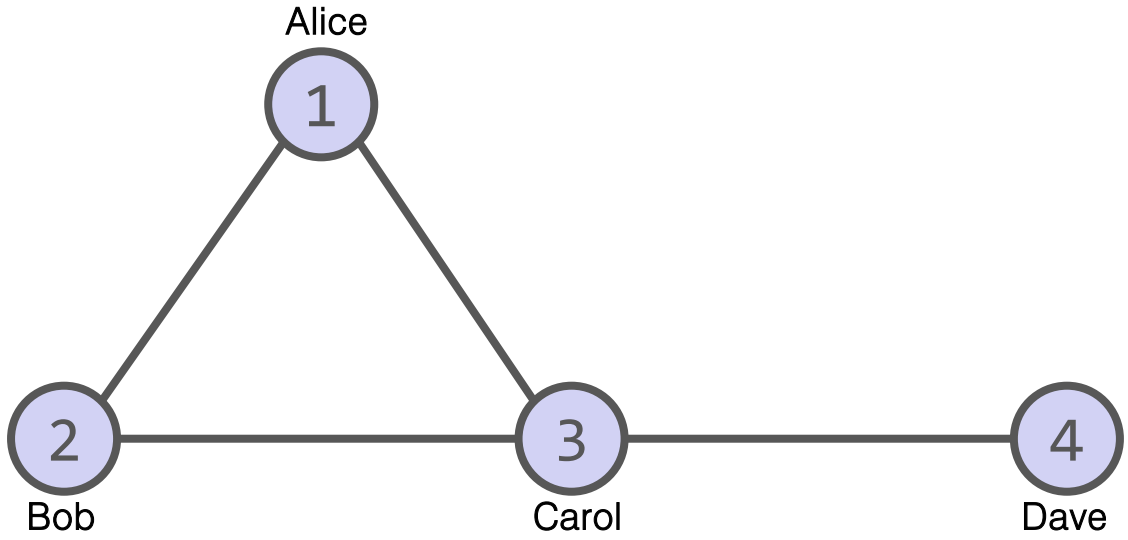
7. Überblick Graph Data Mining

- Einführung
  - Graphmodell
  - Beispiele für Graphdaten
  - Graphbasierte Datenintegration
- Analyse und Mining von Graphdaten
  - Überblick Mining-Verfahren
  - Graph OLAP
  - Pattern Matching
- Frequent Subgraph Mining
  - Problemdefinition
  - Überblick FSM-Algorithmen
  - Effizientes FSM am Beispiel des gSpan-Algorithmus
- Das GRADOOP Framework
  - Beispielsystem

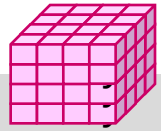


Graphmodell

- Am Beispiel eines sozialen Netzwerks



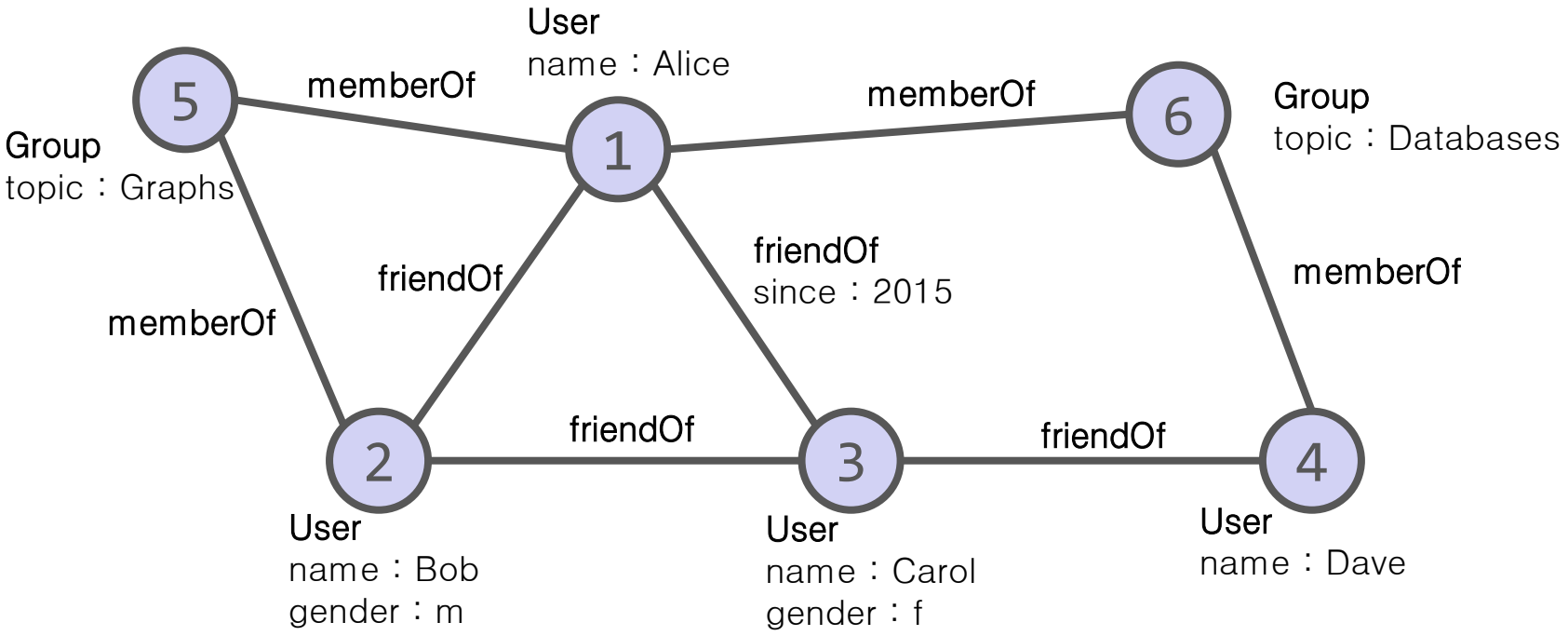
- Knoten (Datenobjekte, z.B. Profile im SN)
- Kanten (Beziehungen, z.B. Freundschaften)



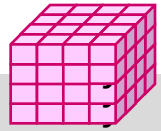
Graphmodell

- Heterogene Graphdaten (z.B. Property Graph Model)

Rodriguez, M.A., Neubauer, P.: Constructions from dots and lines. Bulletin of the American Society for Information Science and Technology 36(6), 35–41 (2010)

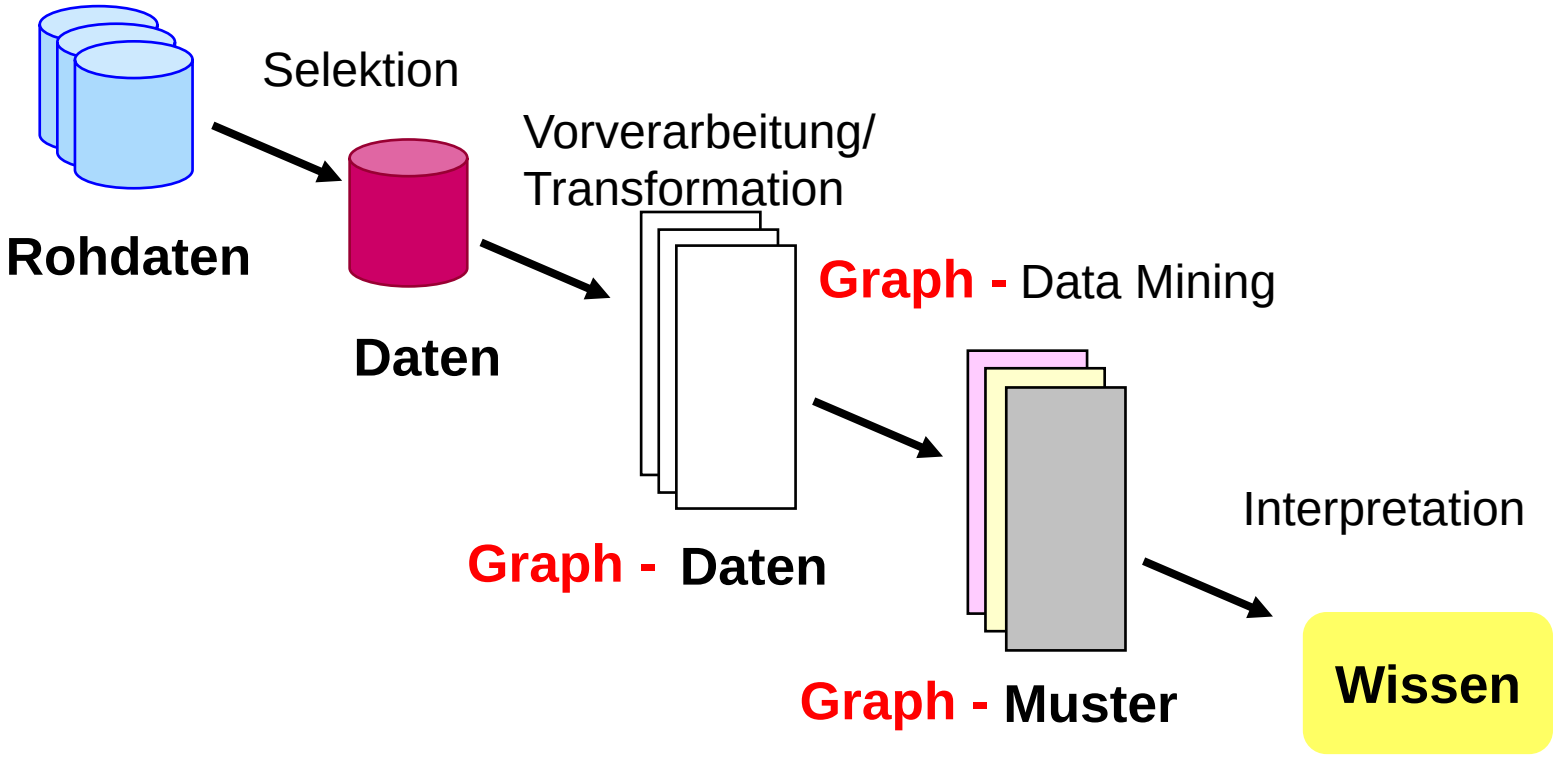


- Labels (Bezeichner, z.B. Beziehungstypen)
- Properties (Eigenschaften, z.B. Nutzername)

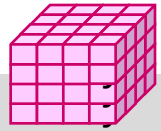


Graph Data Mining

- Prinzip analog zu relationalen / multidimensionalen Daten



- Abweichendes Datenmodell
- Andere Verfahren des Data Mining

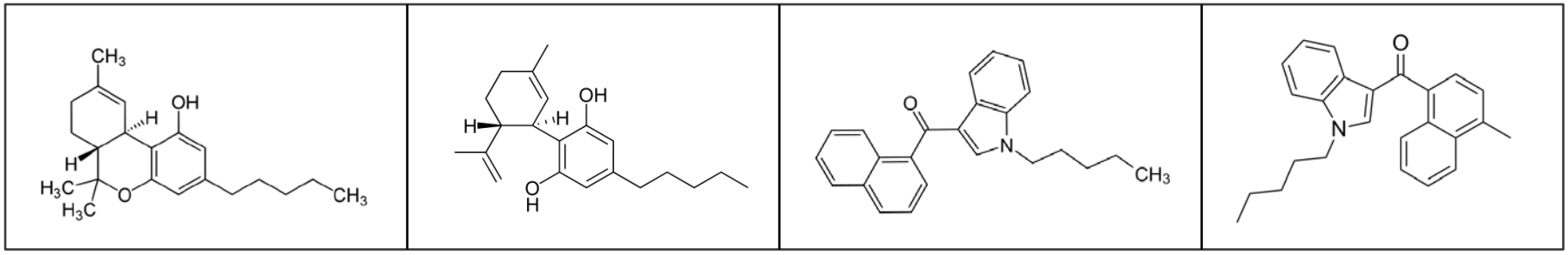


Single Graph vs. Transactional

- Bei vielen Graph Mining-Verfahren wird unterschieden in …

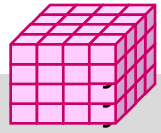
Quelle: Facebook

- … Single Graph Setting (Einzelgraph)

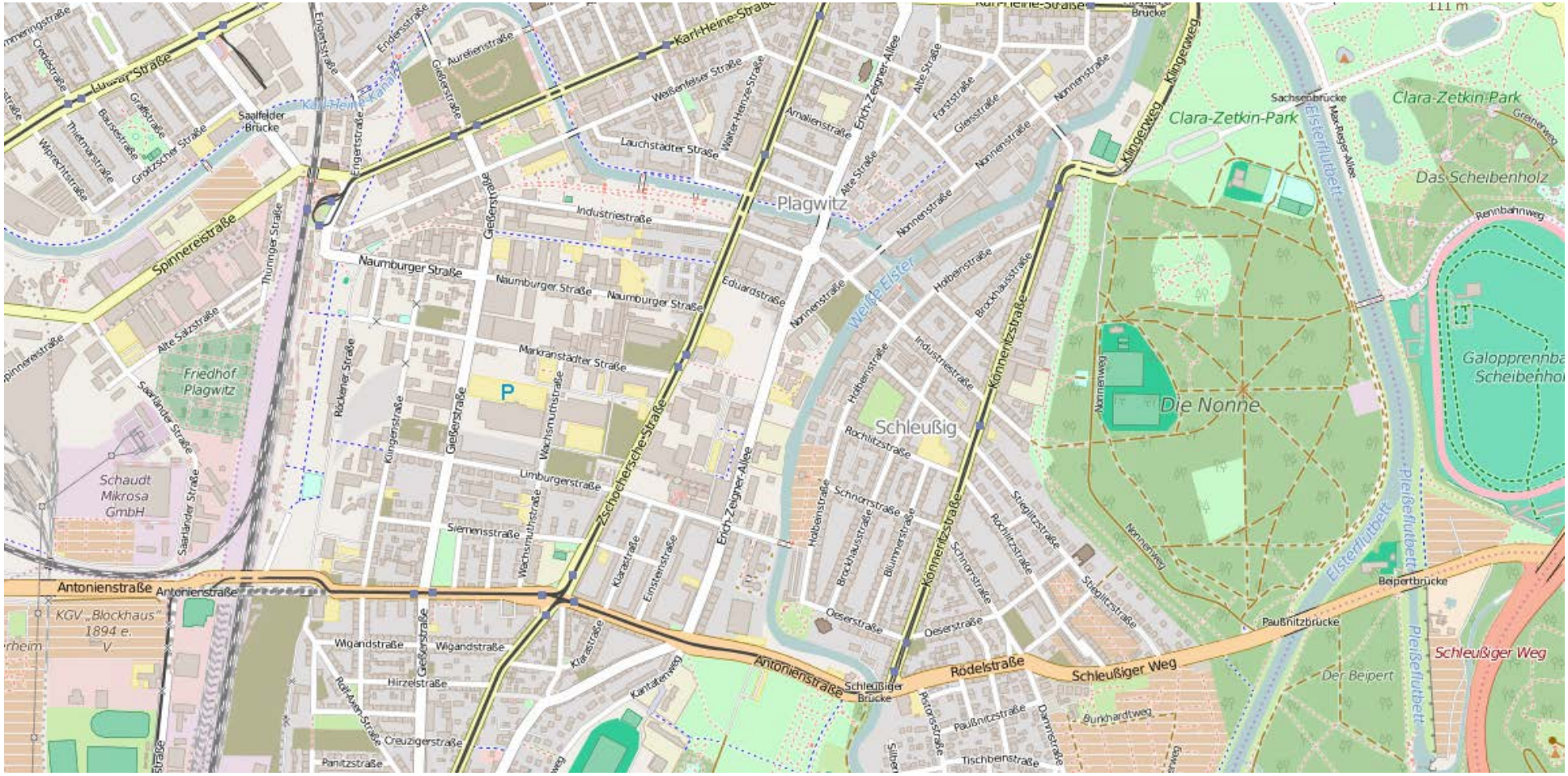


Quelle: Wikipedia

- … Transactional Setting (Graphmenge)

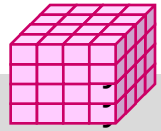


Graphdaten

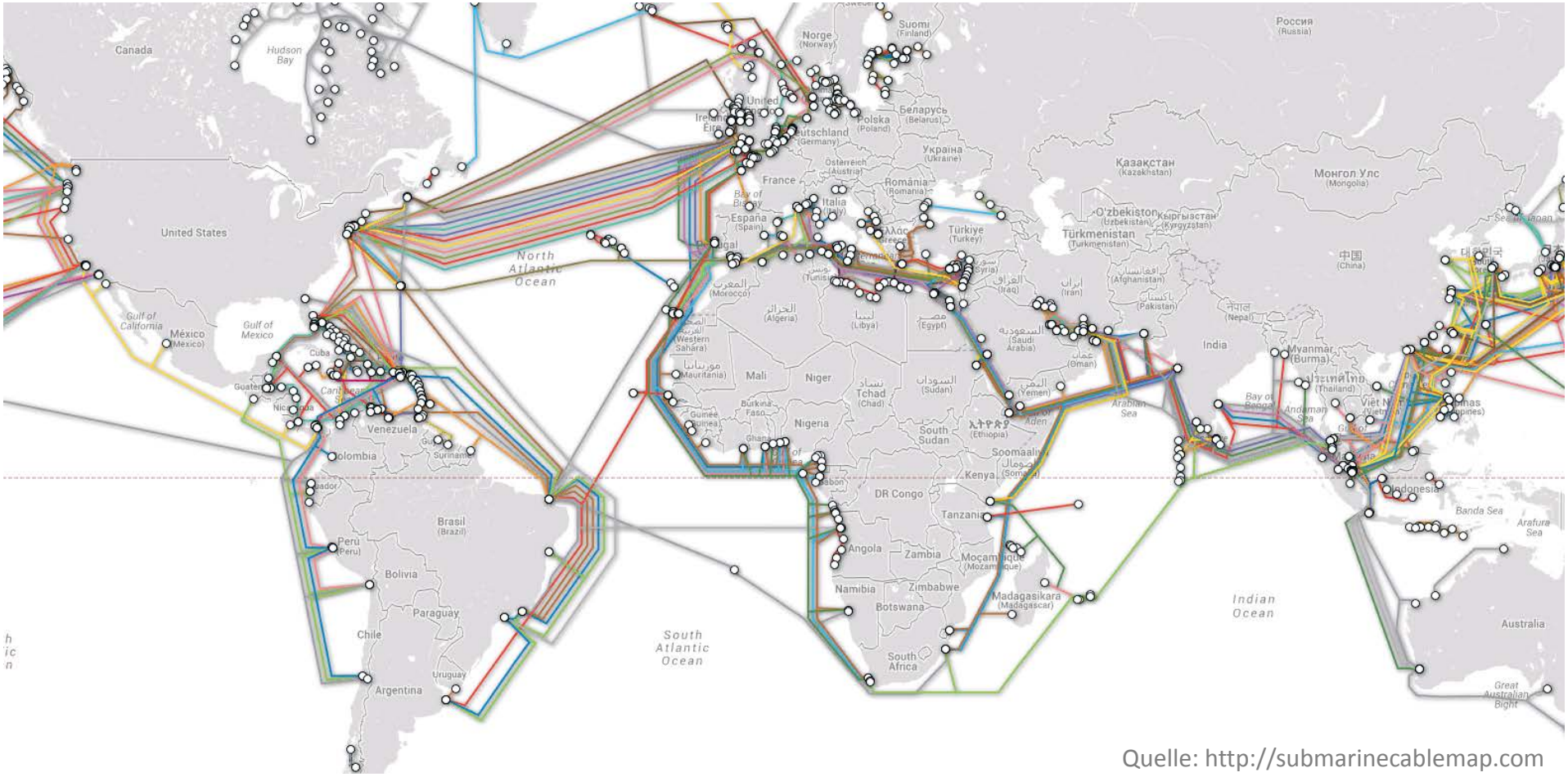


Quelle: OpenStreetMap

- Technologische Netzwerke

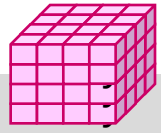


Graphdaten

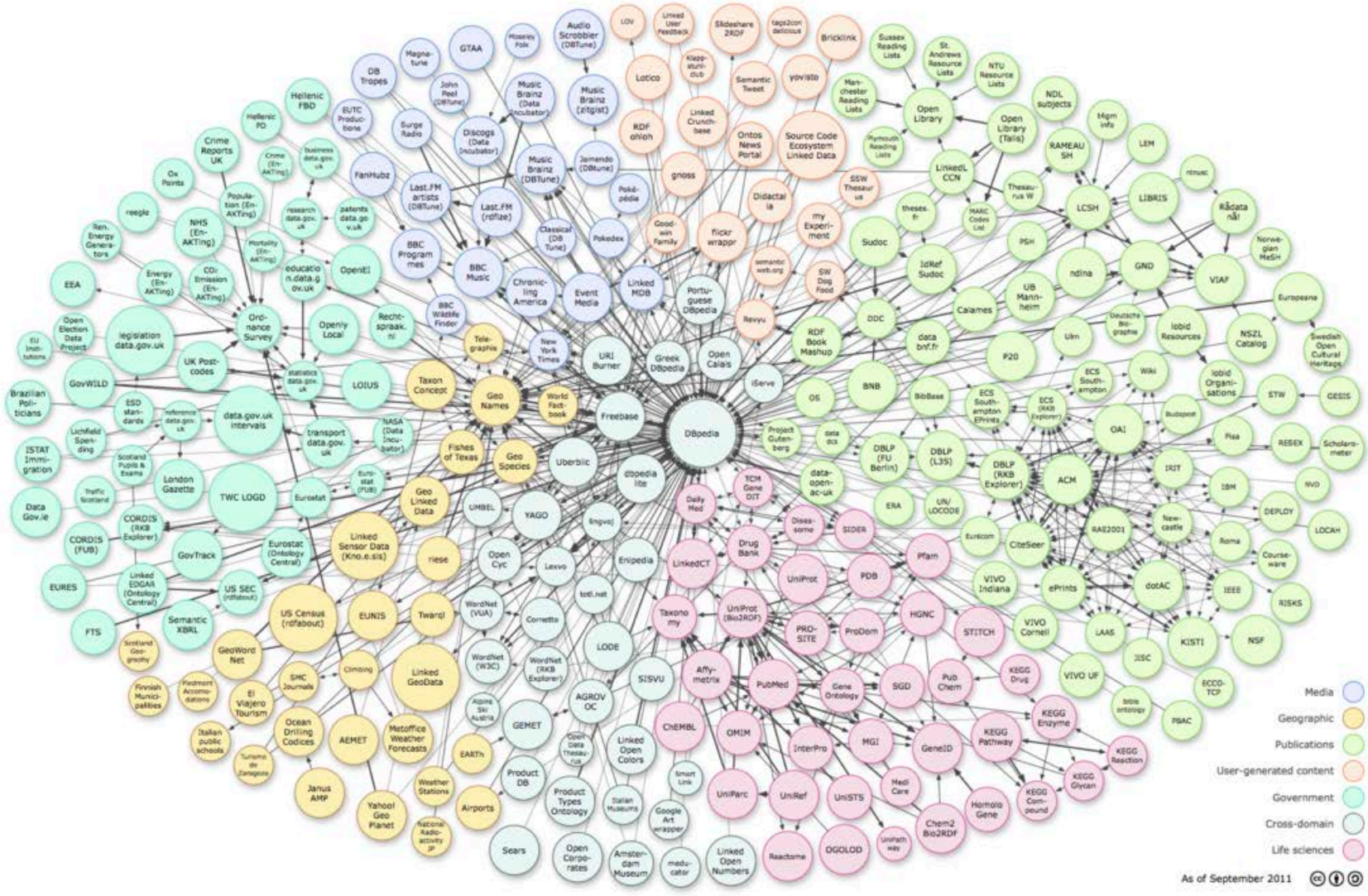


Quelle: http://submarinecablemap.com

- Technologische Netzwerke



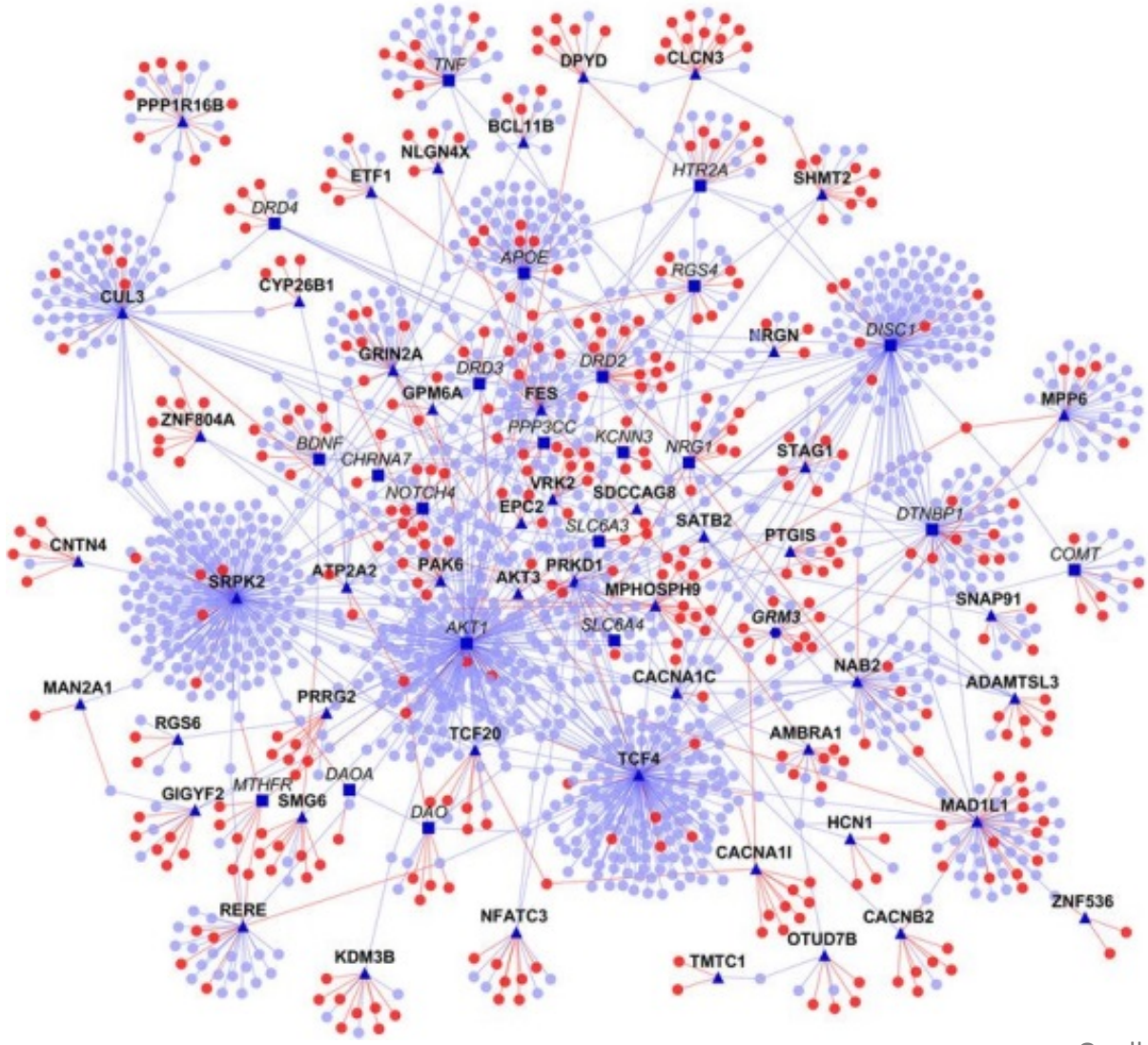
Graphdaten



Quelle: Wikipedia

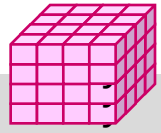
- Informationsnetzwerke

Graphdaten

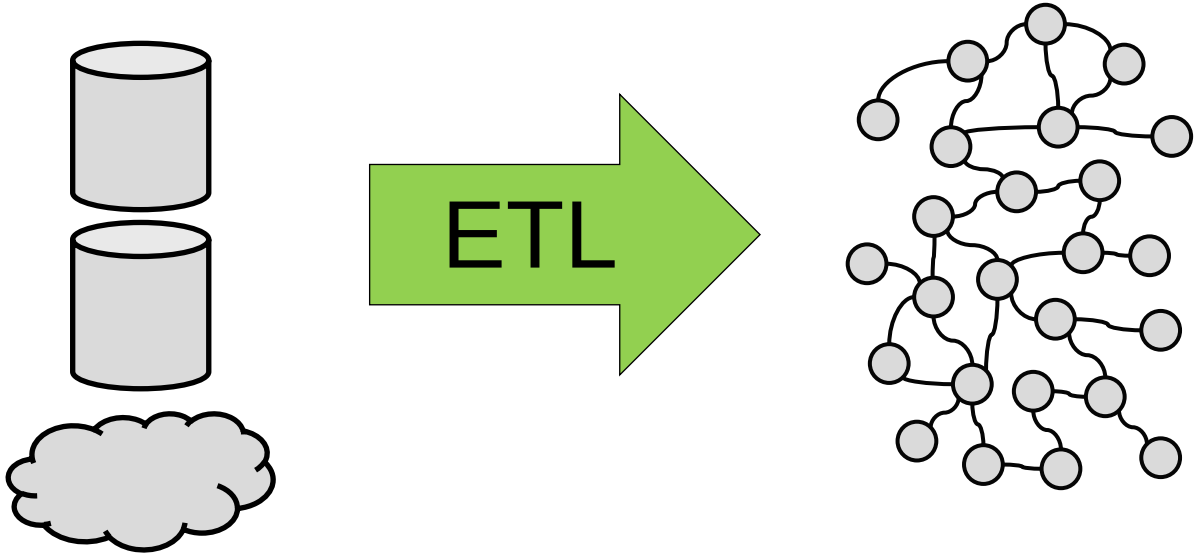


Quelle: Wikipedia

- Biologische Netzwerke



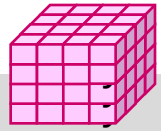
Graphbasierte Datenintegration



- Graphen sind ein Supermodell aller semi-strukturierten und strukturierten Daten

- verschiedene Daten, z.B. relational oder XML können in einen Instanzgraph integriert werden

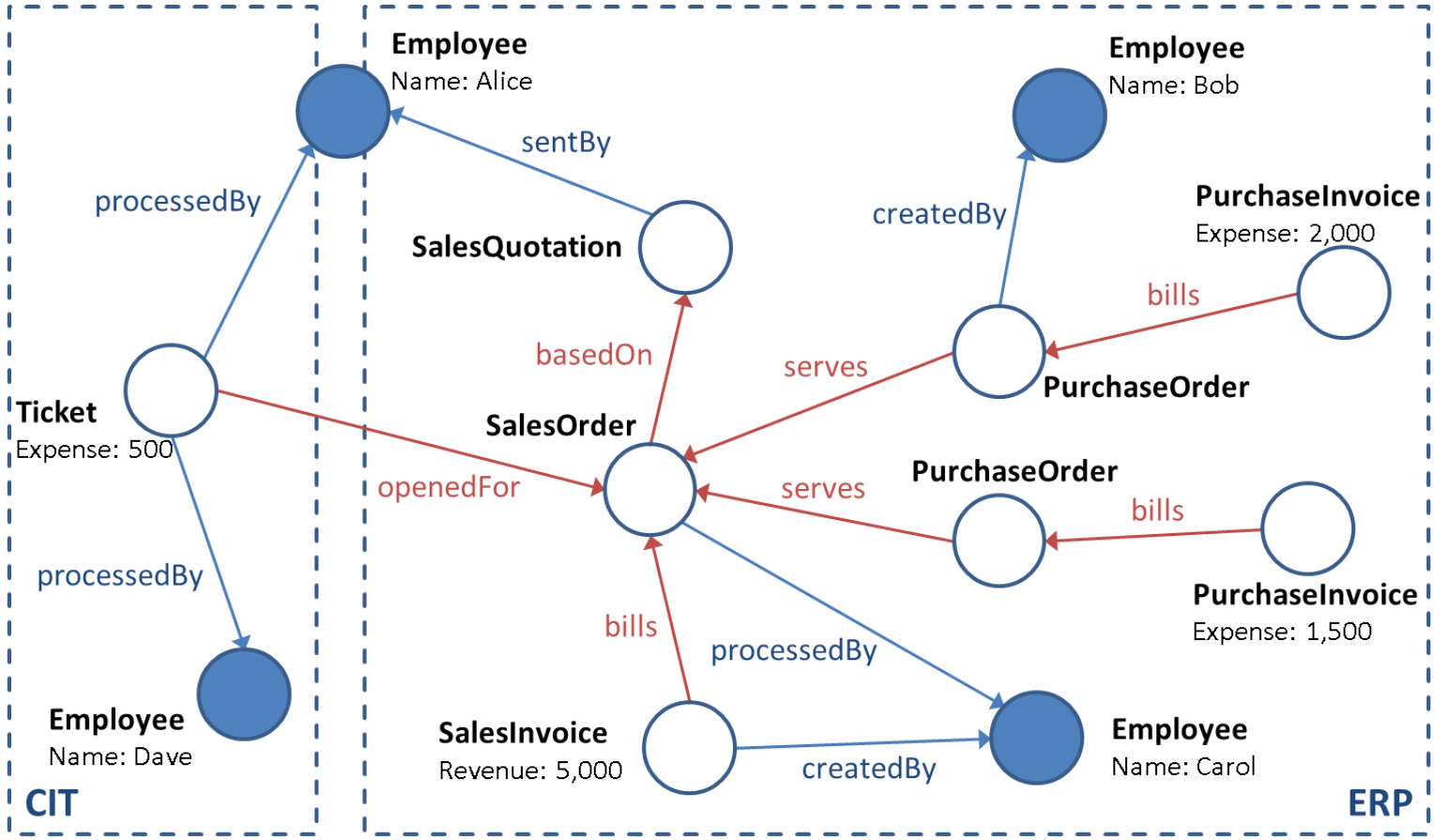
Petermann et al., "BIIIG: enabling business intelligence with integrated instance graphs." Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on. IEEE, 2014.



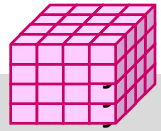
z.B. Relationale Datenbanken → Graph

SalesInvoice			
PK	**Rev**	***SO***	***EMP***
1	5000	321	14
2	-1000	123	13
…			

processedBy	
TicketId	***EMP***
23	13
23	14
…	

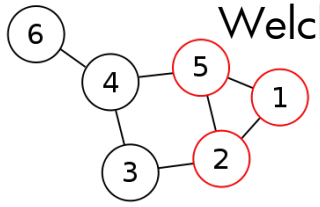


- Datenobjekte → Knoten (z.B. Tabellenzeilen)
- Beziehungen → Kanten (z.B. FK-Spalten, m:n Tabellenzeilen)
- Spalten → Properties (Spaltenname: Wert)

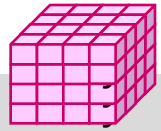


Analyse und Mining-Verfahren

- Connectivity
  - Connected Components — Welche Gruppen von Knoten sind verbunden?
  - Community Detection — Welche Gruppen sind besonders stark verbunden?
  - Page Rank — Welche Knoten sind besonders wichtig?
- Pattern Mining
  - Pattern Matching — Wo tritt ein gegebenes Muster auf?
  - Frequent Subgraph Mining — Welche Muster treten besonders häufig auf?
  - Motif Discovery — Welche signifikanten Muster treten auf?
  - Clique Detection — Finde alle Cliquen!

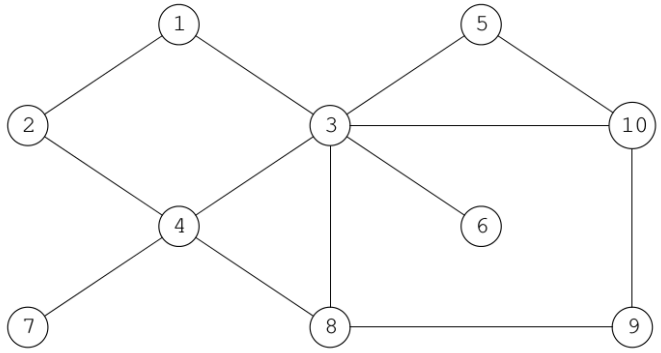


- Graph OLAP
  - Graph Grouping — Gruppiere Freundschaften nach Herkunft der Benutzer!
  - Summarization — Fasse den Graph in 10 Knoten zusammen!
- Weitere Verfahren
  - Clusteranalyse — Analog zu Kapitel 6
  - Klassifizierung — Graphen (Transactional) oder Knoten (Single Graph)



Graph OLAP

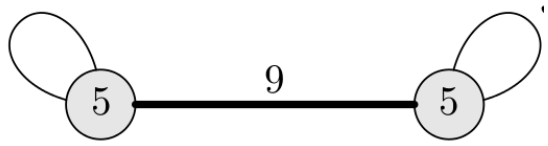
- Zusammenfassen von Graphen



(a) Graph

ID	Gender	Location	Profession	Income
1	Male	CA	Teacher	$70,000
2	Female	WA	Teacher	$65,000
3	Female	CA	Engineer	$80,000
4	Female	NY	Teacher	$90,000
5	Male	IL	Lawyer	$80,000
6	Female	WA	Teacher	$90,000
7	Male	NY	Lawyer	$100,000
8	Male	IL	Engineer	$75,000
9	Female	CA	Lawyer	$120,000
10	Male	IL	Engineer	$95,000

(b) Vertex Attribute Table

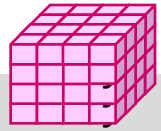


(a) Aggregate Network

Gender	COUNT(*)
Male	5
Female	5

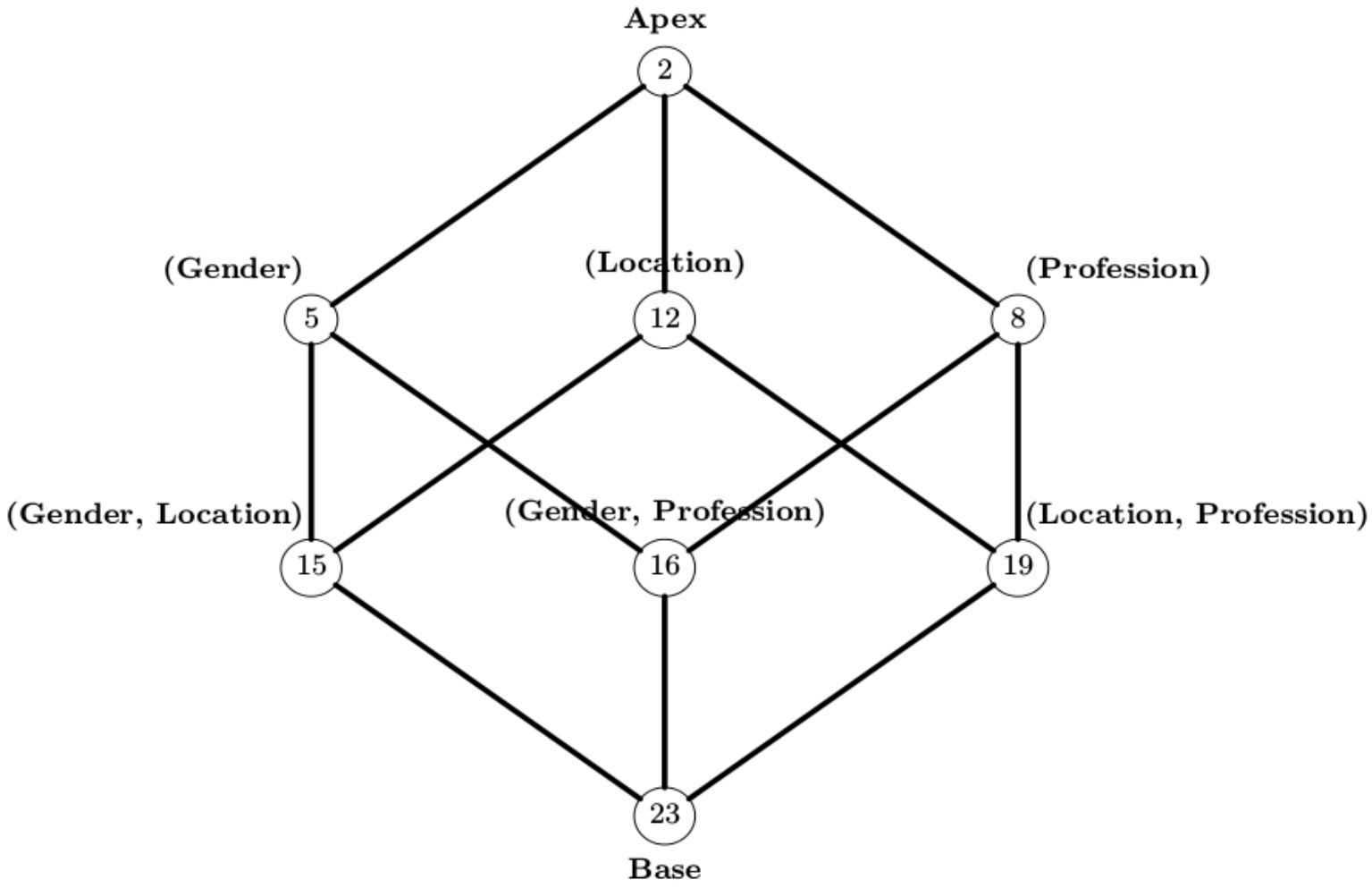
(b) Aggregate Table

Zhao, Peixiang, et al. "Graph cube: on warehousing and OLAP multidimensional networks." Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM, 2011.

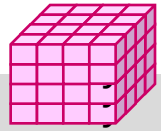


Graph OLAP

- Graph Cube

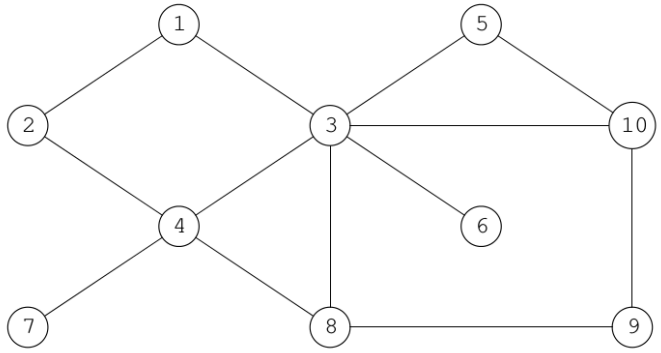


Zhao, Peixiang, et al. "Graph cube: on warehousing and OLAP multidimensional networks." Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM, 2011.



Graph OLAP

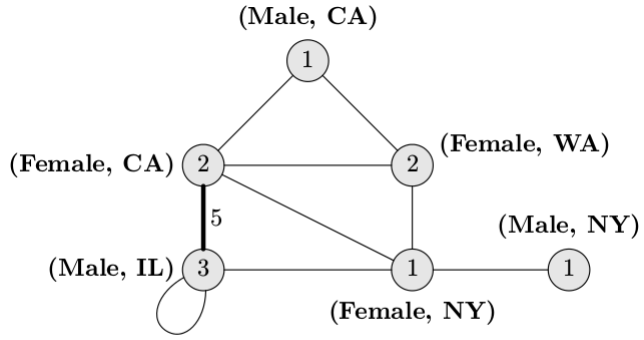
- Multidimensionale Aggregation



(a) Graph

ID	Gender	Location	Profession	Income
1	Male	CA	Teacher	$70,000
2	Female	WA	Teacher	$65,000
3	Female	CA	Engineer	$80,000
4	Female	NY	Teacher	$90,000
5	Male	IL	Lawyer	$80,000
6	Female	WA	Teacher	$90,000
7	Male	NY	Lawyer	$100,000
8	Male	IL	Engineer	$75,000
9	Female	CA	Lawyer	$120,000
10	Male	IL	Engineer	$95,000

(b) Vertex Attribute Table

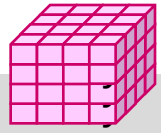


(a) Aggregate Network

Gender	Location	COUNT(*)
Male	CA	1
Female	CA	2
Female	WA	2
Male	IL	3
Male	NY	1
Female	NY	1

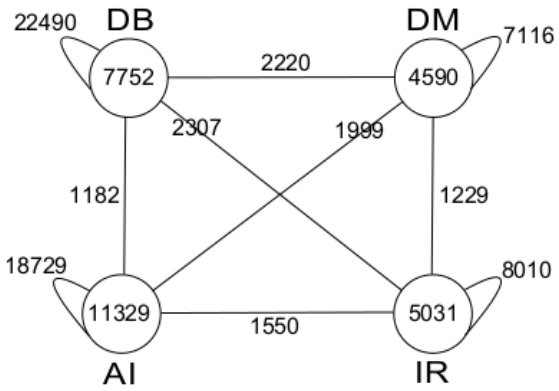
(b) Aggregate Table

Zhao, Peixiang, et al. "Graph cube: on warehousing and OLAP multidimensional networks." Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM, 2011.

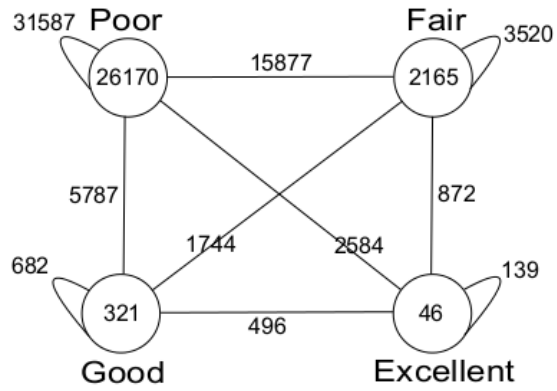


Graph OLAP

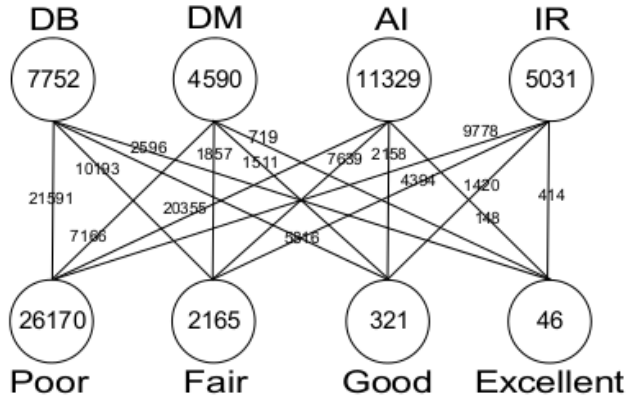
- Multidimensionale Aggregation



(a) (Area)

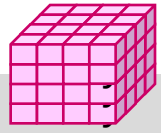


(b) (Productivity)



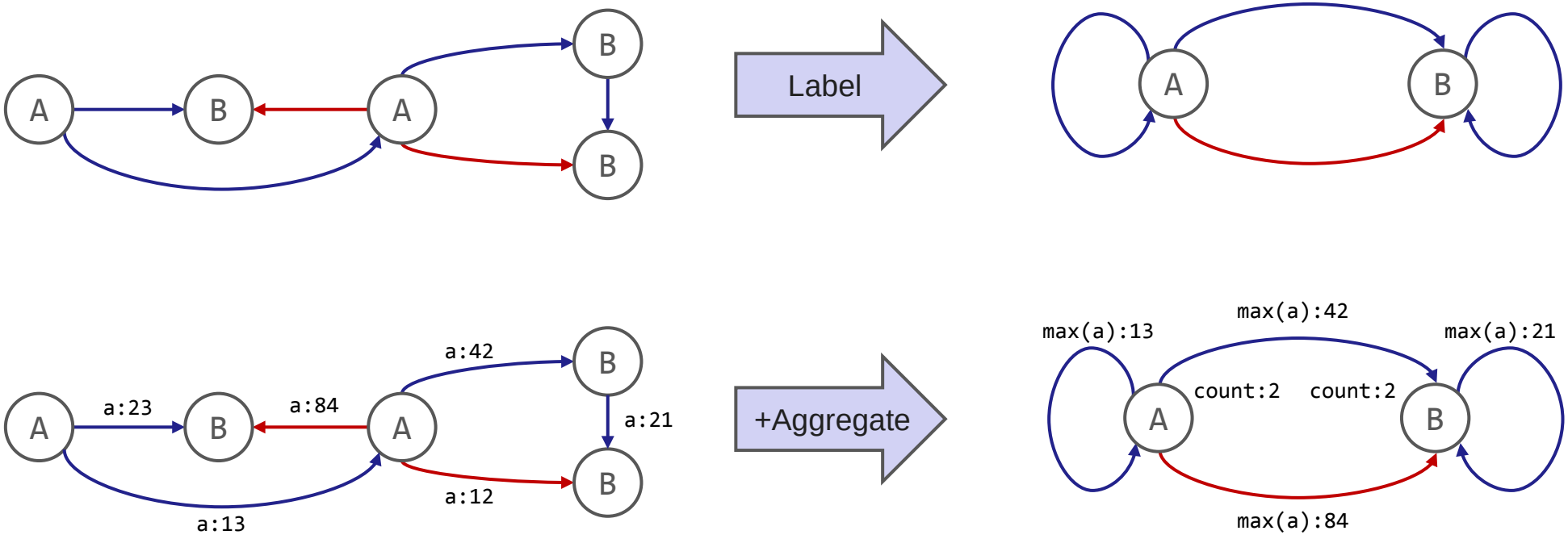
(a) Area ⋈ Productivity

Zhao, Peixiang, et al. "Graph cube: on warehousing and OLAP multidimensional networks." Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM, 2011.

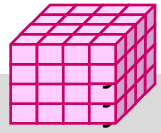


Graph OLAP

- Grouping-Queries
- „Zusammenfassungen On-Demand“

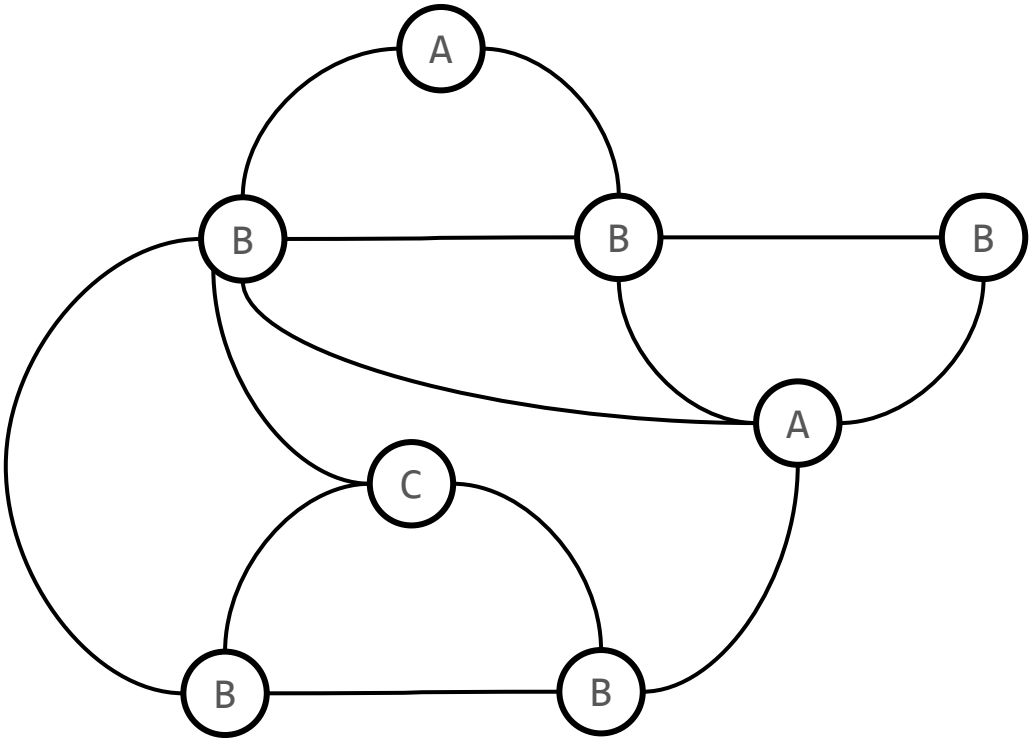


Junghanns et al., Analyzing Extended Property Graphs with Apache Flink
Proc. Int. SIGMOD workshop on Network Data Analytics (NDA) 2016-07

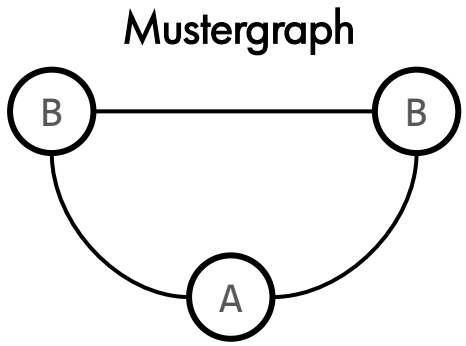


Pattern Matching

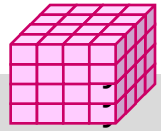
- Auffinden von vorgegebenen Mustern in Graphdaten



Datengraph

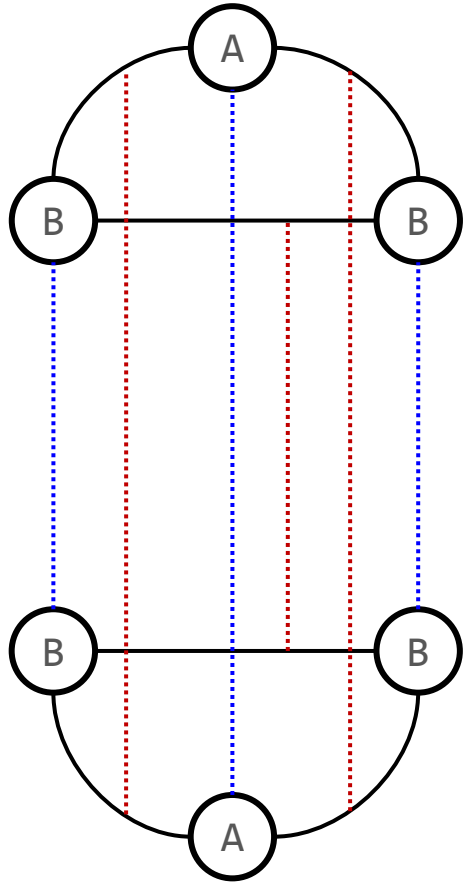


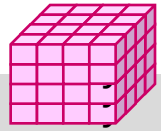
Mustergraph



Exkurs Isomorphie

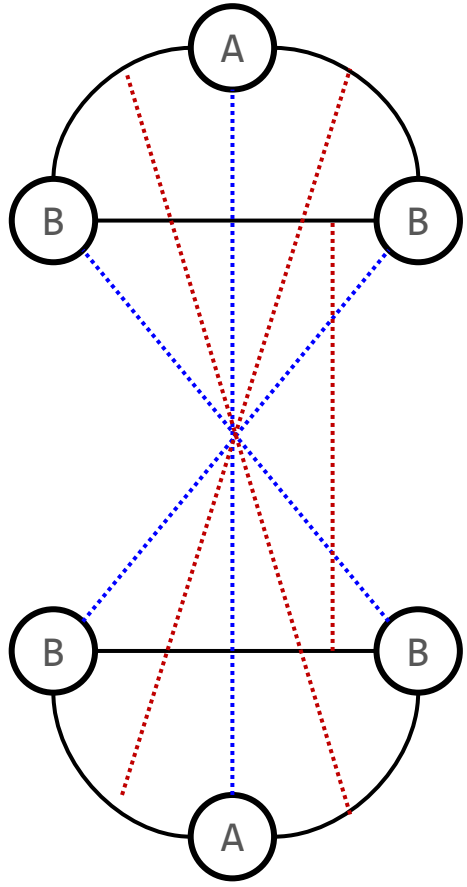
- Exakte Übereinstimmung von Graphen
- Knotenabbildung $v : V(G) \leftrightarrow V(H)$
- Kantenabbildung $e : E(G) \leftrightarrow E(H)$





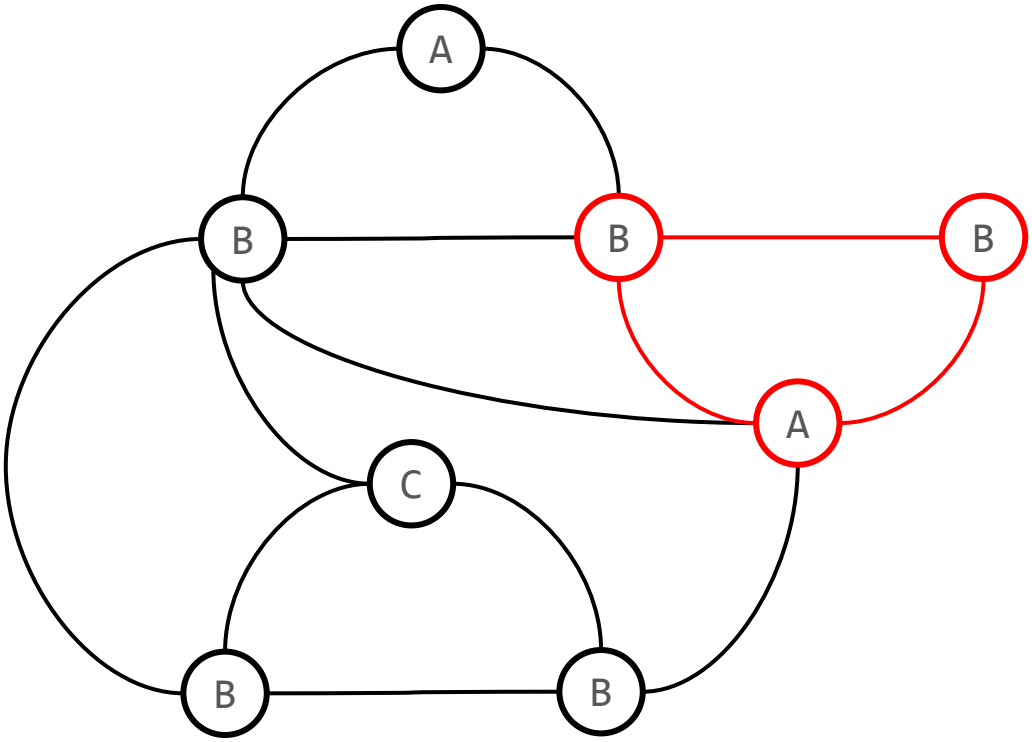
Exkurs Isomorphie

- Exakte Übereinstimmung von Graphen
- Knotenabbildung $v : V(G) \leftrightarrow V(H)$
- Kantenabbildung $e : E(G) \leftrightarrow E(H)$
- Mehrfache Varianten möglich (Automorphie)

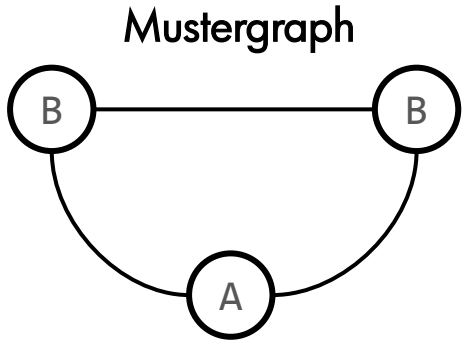


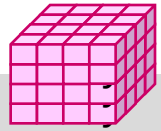
Pattern Matching

- Teilgraph-Isomorphie



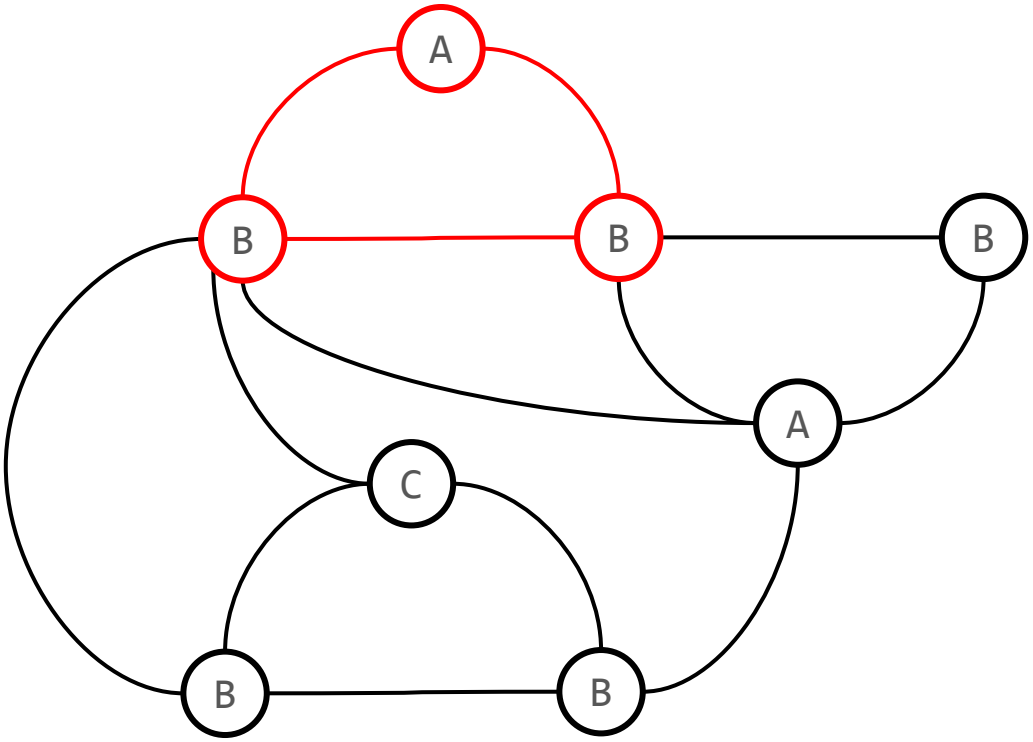
Datengraph



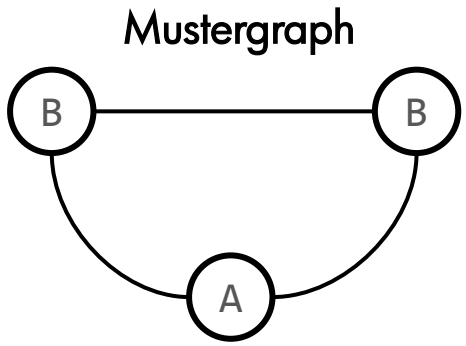


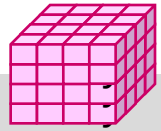
Pattern Matching

- Teilgraph-Isomorphie



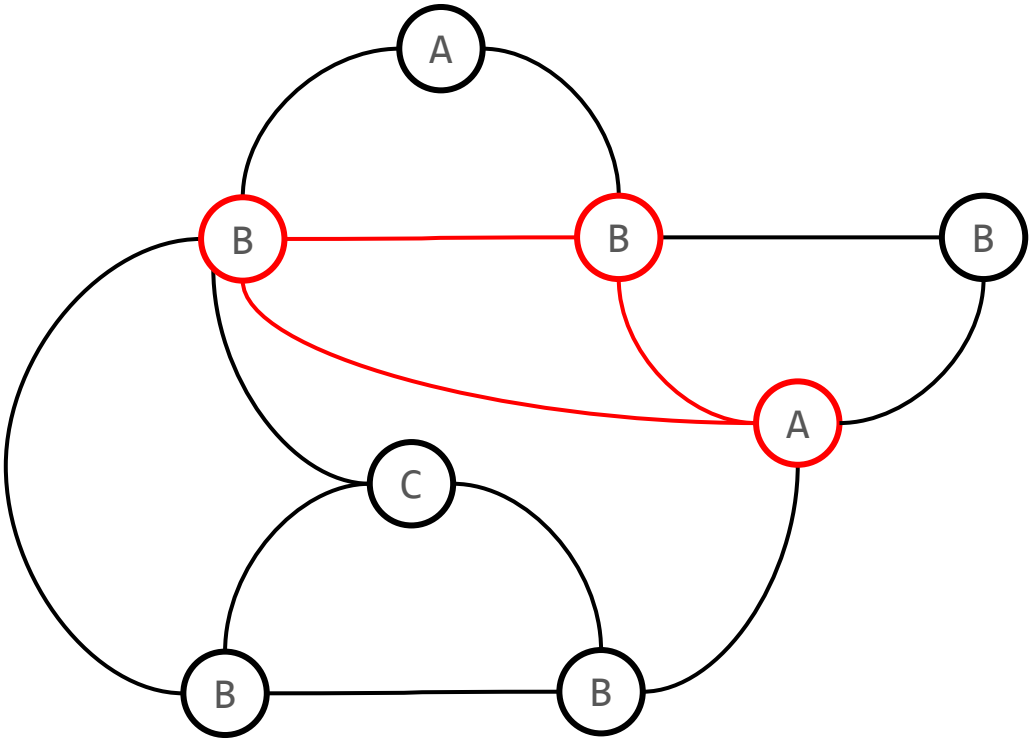
Datengraph



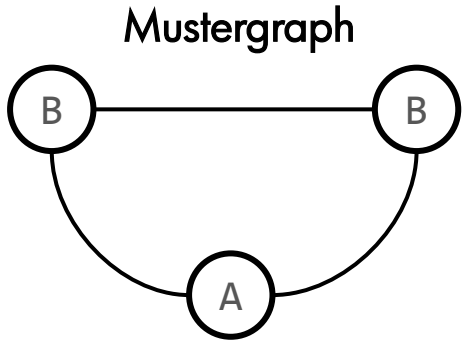


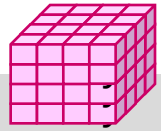
Pattern Matching

- Teilgraph-Isomorphie



Datengraph



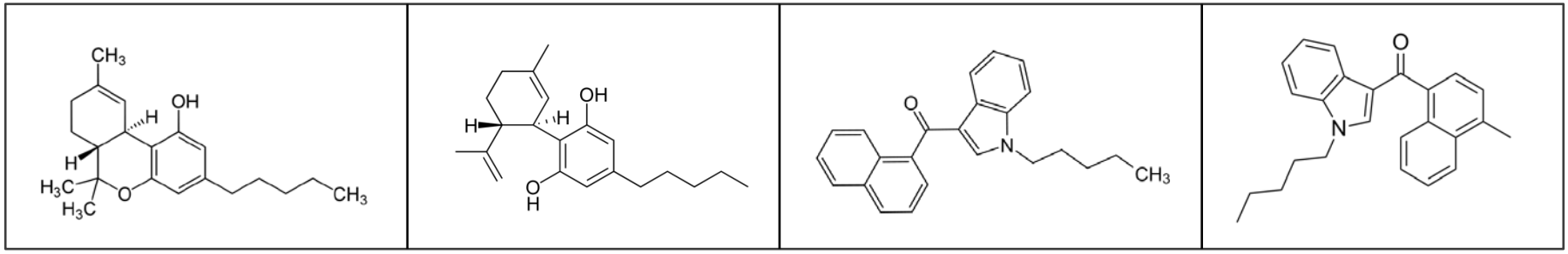


Graph Pattern Mining

- ist das Finden interessanter Teilgraphen in

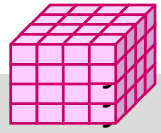
Quelle: Facebook

- einem Einzelgraph (Single Graph Setting) oder



Quelle: Wikipedia

- einer Graphmenge (Transactional Setting)

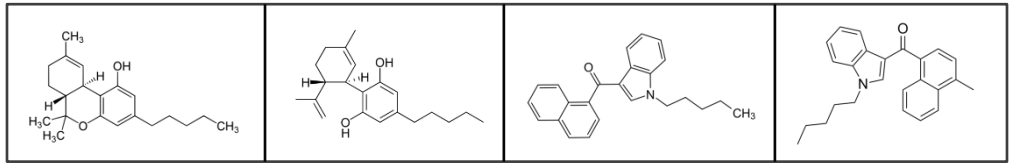


Frequent Subgraph Mining

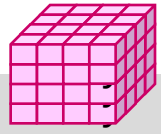
- Interessantheit → Häufigkeit

- Single Graph Setting:
  - $g = (V, E)$
  - Mindesthäufigkeit f_{min}
  - Häufigkeit: $f(s, g) \mid s \subseteq g$
  - Ein Teilgraph ist interessant, wenn $f(s, g) \geq f_{min}$

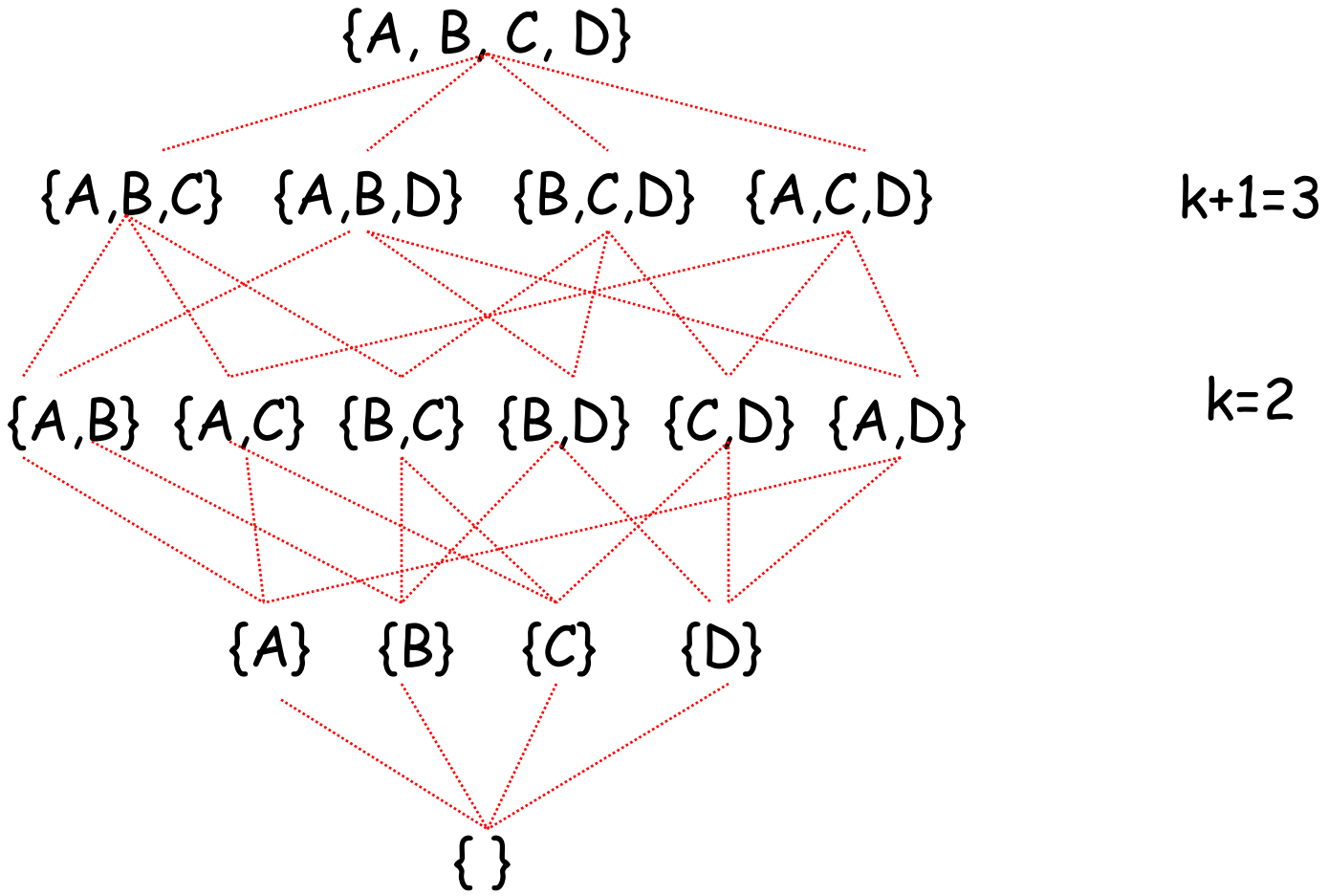
- Transactional Setting



  - $G = \{g_0, g_1, .., g_n\}$
  - Minimum Support sup_{min}
  - Support: $sup(s, G) = {}^{|G_{sup}|}/_{|G|} \mid \forall\, g \in G_{sup} : s \subseteq g$
  - Ein Teilgraph ist interessant, wenn $sup(s, G) \geq sup_{min}$

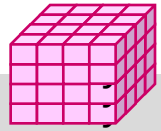


Verbandstruktur von Itemsets

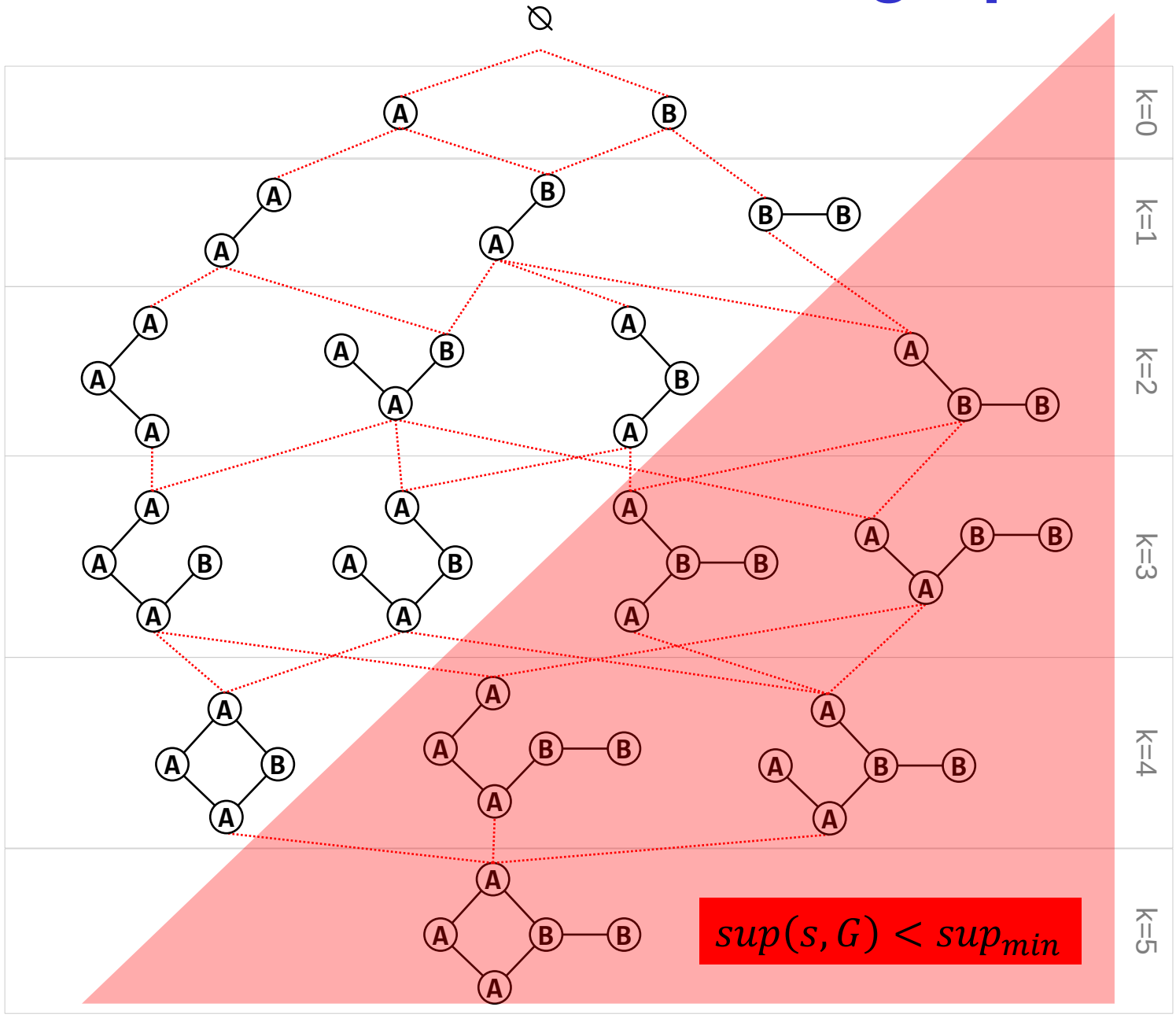


Frequent Subgraph Mining

- Itemset Mining
  - Übereinstimmende Elemente
- Sequence Mining
  - Übereinstimmende Elemente und Reihenfolge
- Tree Mining
  - Übereinstimmende Elemente und Struktur
- Subgraph Mining
  - Übereinstimmende Elemente und Struktur mit Zyklen



Verbandstruktur von Teilgraphen



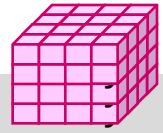
Frequent Subgraph Mining

- A-Priori
  - Finde alle häufigen Teilgraphen mit einer Kante
  - Finde alle häufigen Teilgraphen mit zwei Kanten
  - Bis alle häufigen Teilgraphen mit k Kanten entdeckt:
    - Bilde Kandidaten durch Fusion häufiger Teilgraphen
    - Ermittle häufige Teilgraphen per Teilgraph-Isomorphie-Test

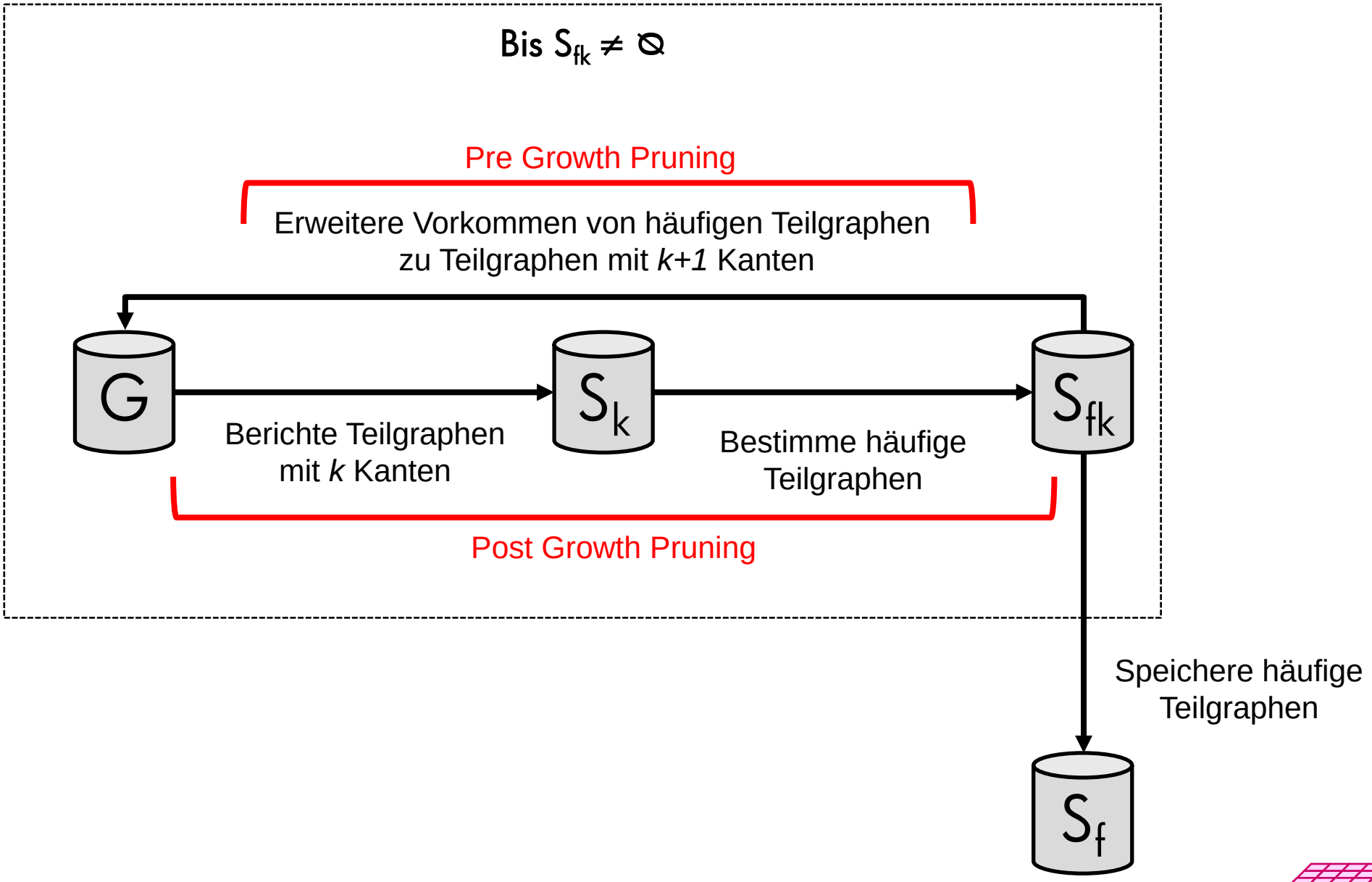
Kuramochi, Michihiro, and George Karypis. "Frequent subgraph discovery." ICDM 2001

- Pattern Growth
  - Beginne mit Teilgraphen mit einer Kante
  - Bis alle häufigen Teilgraphen mit k Kanten entdeckt:
    - Sammle alle unterstützten Teilgraphen mit k-Kanten
    - Ermittle häufige Teilgraphen durch Gruppierung nach Normalform
    - Erweitere alle Vorkommen um eine Kante

Yan, Xifeng, and Jiawei Han. "gspan: Graph-based substructure pattern mining." ICDM 2002



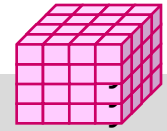
Pattern Growth



gSpan Algorithmus

- Pattern Growth
- Sehr effektiv durch Kombination verschiedener Strategien
- Vermeiden von Teilgraph-Isomorphie-Tests (NP-vollständig)
- Aggressive Einschränkung des Suchraums
- Systematisches Wachstum von Teilgraphen

Yan, Xifeng, and Jiawei Han. "gspan: Graph-based substructure pattern mining." ICDM 2002



gSpan Algorithmus

- Ein Teilgraph kann nur häufig sein,
  wenn auch sein Parent im Verband häufig ist.

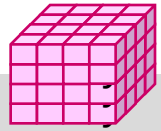
- Häufige Teilgraphen müssen häufige Labels haben

- Vorverarbeitung:

  – Ermittlung häufiger Labels von Knoten und Kanten

  – Dictionary Encoding (Häufiges Label -> Kleine Ganzzahl)

  – Knoten und Kanten ohne häufiges Label entfernen

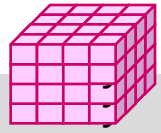


gSpan Algorithmus

- Normalform statt Teilgraph-Isomorphie-Test

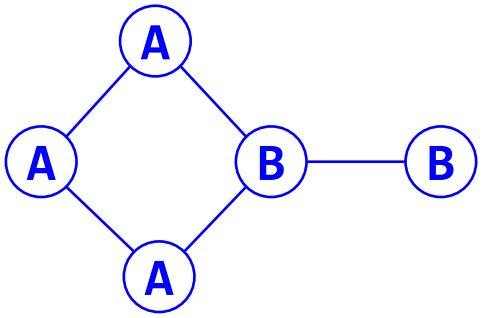
- Normalform
  - Jeder (zusammenhängende) Graph lässt sich durch verschiedene Tiefensuchen beschreiben
  - Der Verlauf der Tiefensuche lässt sich kodieren (DFS Code)
  - Es gibt eine lexikografische Ordnung von DFS Codes
  - Es existiert für jeden (zusammenhängende) Graphen ein minimaler DFS Code
  - ➢ Minimaler DFS Code als Normalform

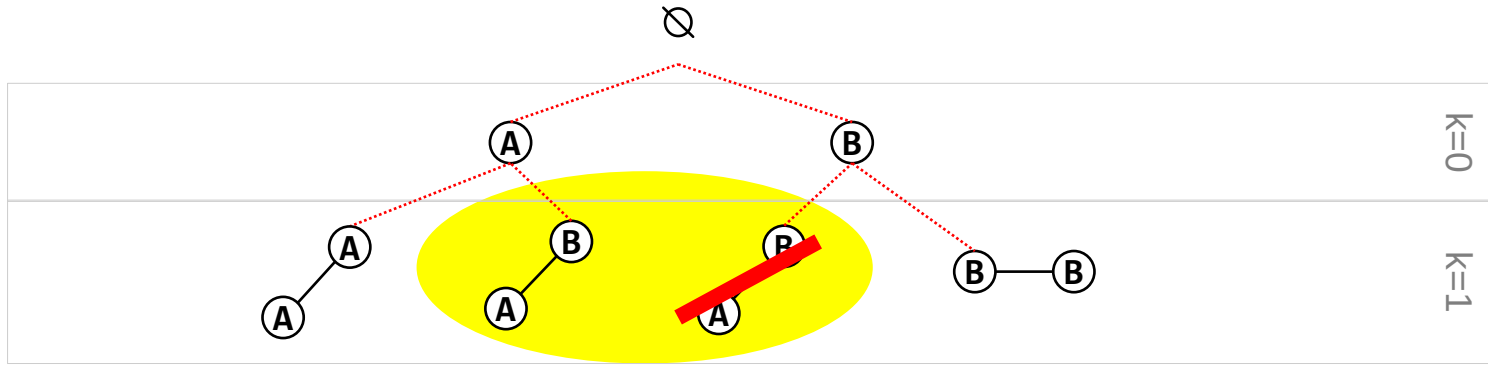
- Alle Teilgraphen werden in Normalform gezählt
  - ➢ Kein Isomorphie-Test (Abbildungen ermitteln) notwendig



gSpan Algorithmus

- Teilgraph-Baum

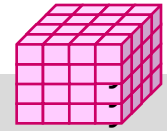




- Problem: Duplikate

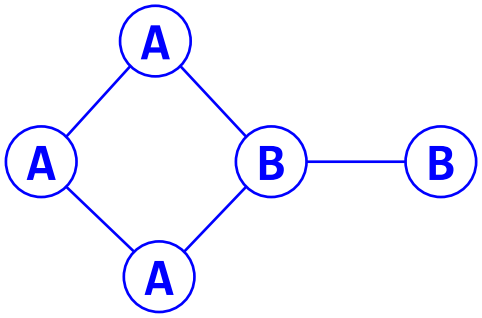
- Lösung 1: Reihenfolge (Lexikografische Ordnung) zählt (Post)

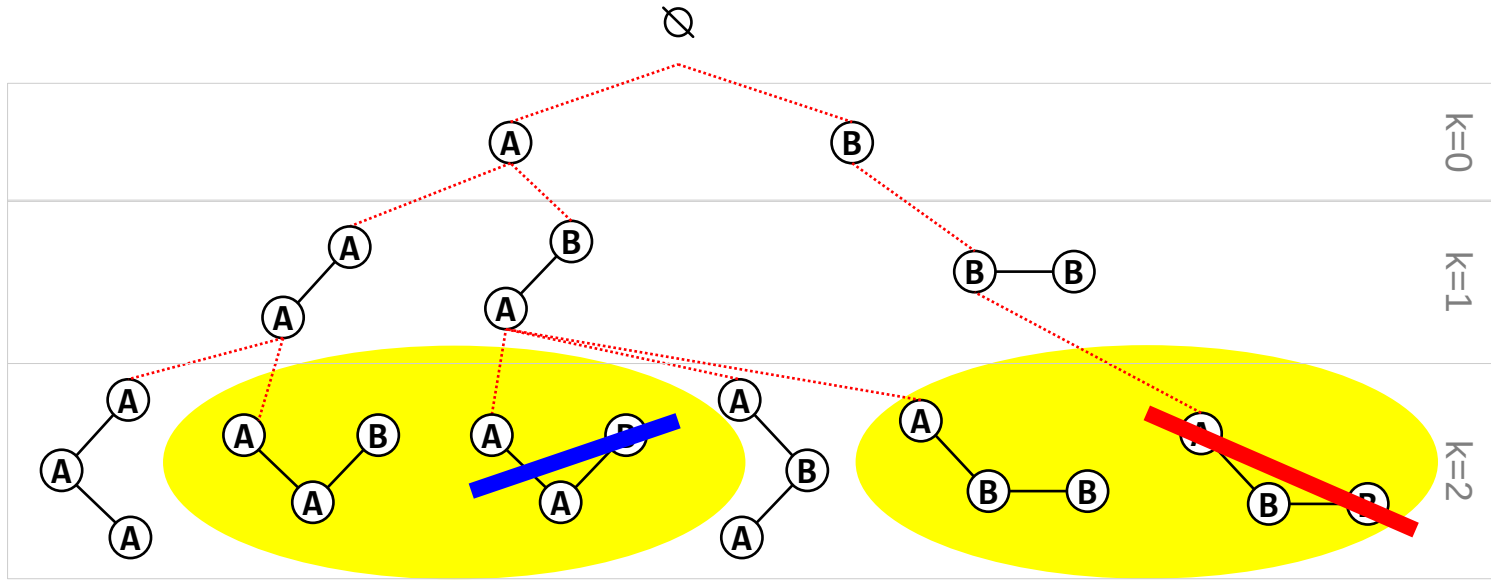
- (0:A)-e-(1:B) < (0:B)-e-(1:A)
- (0:B)-e-(1:A) ist nicht minimal



gSpan Algorithmus

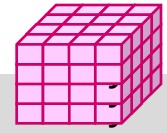
- Teilgraph-Baum





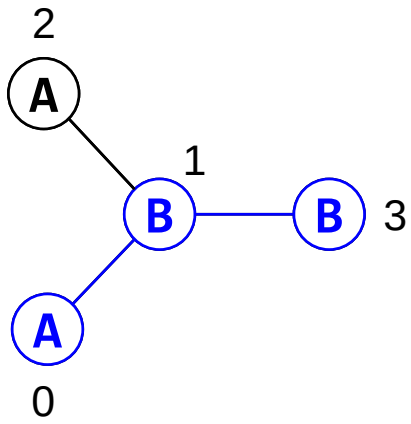
- Lösung 2: Kleinstes Label einer Erweiterung darf nicht kleiner sein als das des 0-edge Parents (Pre)

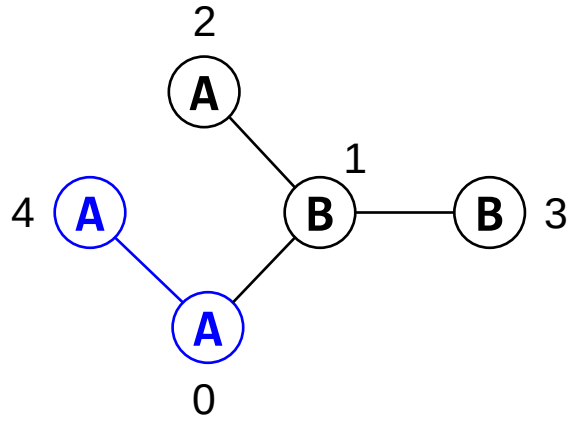
- Wieder 1 : (0:A)-e-(1:A),(1:A)-e-(2:B) < (0:A)-e-(1:B),(0:A)-e-(1:A)



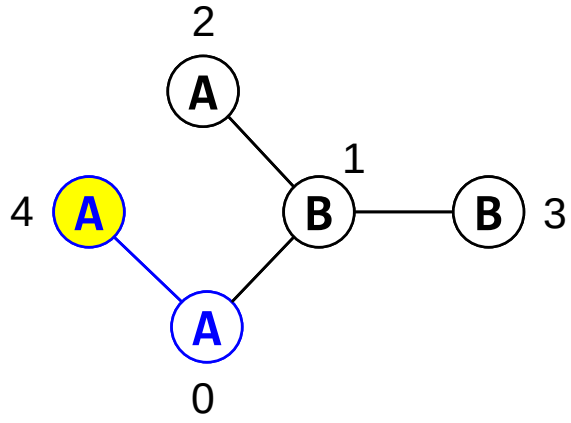
gSpan Algorithmus

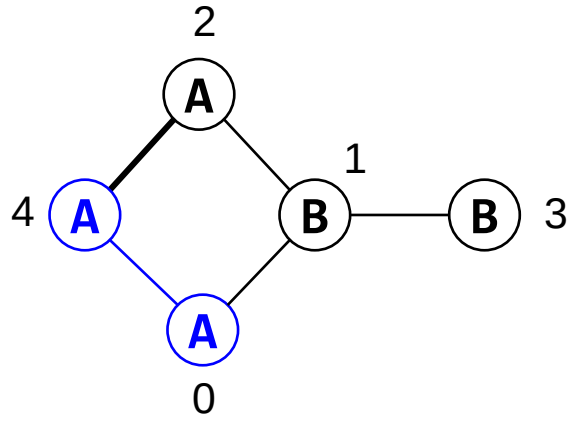
- Lösung 3: Erweiterung nur vom ganz rechten Pfad (Pre)

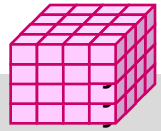




- Lösung 4: Zyklenschluss nur vom ganz rechten Knoten (Pre)

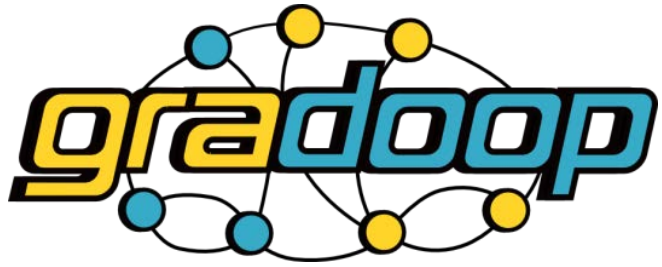






GRADOOP-Framework

- Open Source System zur Graphanalyse
- Entwickelt an der Abteilung Datenbanken
- Beinhaltet verschiedene analytische Algorithmen, z.B.:
  - Pattern Matching
  - Graph Grouping
  - Frequent Subgraph Mining



www.gradoop.com

- Erweitertes Property Graph Model
  - Unterstützung für Single Graph und Transactional Setting
- 3rd Pary Library für Apache Flink