Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

Graph Machine Learning

Themenantrag Master Thesis

▸ Q1/22, Thomas Iten

Bild: https://cutewallpaper.org

Graph Machine Learing

Explorative Arbeit im Bereich Link Prediction

Berner Fachhochschule | Haute école spécialisée bernoise | Bern University of Applied Sciences

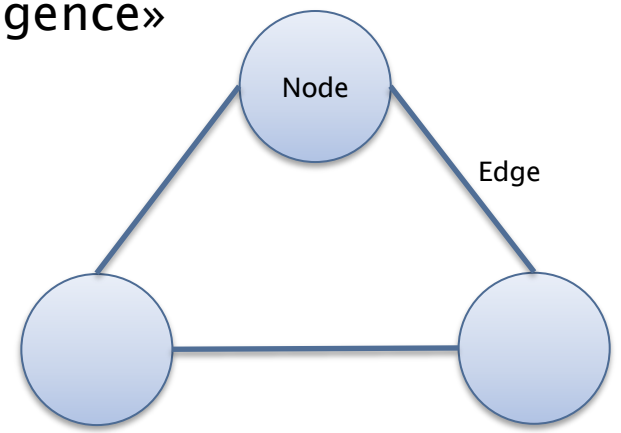
Eckdaten

Angaben zu Autor	
Autor/Autorin:	Thomas Iten
Ausbildung/ Höchster Abschluss:	Diplom Ingenieur FH Elektrotechnik Nachdiplom Softwareingenieur HTL-NDS Nachdiplom Wirtschaftsingenieur FH STV
Firma/Ort:	Die Mobiliar, Bern
Studiengang:	MAS Data Science
Eingangskompetenzen	
1. CAS/Semester:	CAS Datenanalyse / HS19
2. CAS/Semester:	CAS Practical Machine Learning / FS20
3. CAS/Semester:	CAS Data Visualization / HS20
4. CAS/Semester:	CAS Artificial Intelligence / HS21

Angaben zur Master-Thesis	
Semester:	FS22
Themensponsor / Firma:	Die Mobiliar, Bern
Themensponsor / Betreuungsperson:	Matthias Brändle Verantwortlicher Datenstrategie
Art der Arbeit / Referenzrahmen:	Explorative Arbeit

Ausgangslage

- Mit **Graphen** lassen sich **komplexe Strukturen** wie physikalische Systeme, Molekularstrukturen, Verkehrsnetze oder soziale Netzwerke abbilden.
- Neben der Graph **Struktur** beinhalten die **Knoten** und **Verbindungslinien** der Graphen ebenfalls wertvolle **Eigenschaften**.
- Für die **Verarbeitung** solcher Informationen kommen unter anderen **Machine Learning** (ML) Algorithmen zum Einsatz.
- Diese sind in der Lage Informationsmerkmale mit Graph Strukturen zu kombinieren, und davon zu lernen und **Vorhersagen** (auf Graph-, Node- oder Edgelevel) zu treffen.
- Mit den beiden CAS «Practical Machine Learning» und «Artificial Intelligence» konnte ich erste **Erfahrungen mit Machine Learning** sammeln.
- Die **Kombination** von Graphen mit ML ist spannend und **fasziniert mich persönlich** sehr.
- Gerne würde ich das Thema im Rahmen der Master-Thesis intensiver erkunden und praktische Erfahrungen sammeln.



Ziele

- Mit der Master-Thesis soll das **Wissen über Graphen in Kombination mit dem Einsatz von ML** Algorithmen aufgebaut und mit praktischen Experimenten nachvollzogen werden.
  - Dazu werden öffentliche Daten ausgewählt und eingesetzt, so dass die Erkenntnisse und Experimente für alle Interessierten im Internet frei zugänglich sind.
  - Ziel ist es, mit dem so erlangten Wissen abschätzen zu können, für welche Geschäftsanwendungen die Technologie in Zukunft nutzbringend eingesetzt werden kann.
- Darauf aufbauend soll eine **Fragestellung im Kontext der Mobiliar Versicherung** vertieft untersucht werden.
  - Mit dem «Enterprise Data Catalog» (EDC) stellt die Mobiliar eine **Graph** basierte Anwendung zur Verfügung.
  - Diese importiert die **Metadaten** von verschieden Datenquellen (Datenbanken, Protokollen, Mitarbeiter Stammdaten, etc.) und führt die Datensätze in einem Graphen zusammen.
  - Eine Aufgabe ist dabei die **Verknüpfung** von Teilnehmernamen der Sitzungsprotokolle mit den effektiven Personen aus den Stammdaten.

Ziele II

- Die Vorhersage von fehlenden Verbindungen (**Link Prediction**) ist eine häufige Aufgabe beim Arbeiten mit Graphen.
- Die zu untersuchenden Fragestellung lautet also: **Wie können mit Link Prediction Techniken die Daten von verschiedenen Datenquellen effizient miteinander verknüpft werden?**
- Konkret soll dazu in einer Fallstudie die erwähnte Verknüpfungsproblematik der EDC Anwendung mit diversen Link Prediction Techniken (wie Ressource Allocation, Jaccard Coefficient, Common Neigbor oder Supervised Classification) untersucht und verglichen werden.

Zu klärende Fragen

- Wie kann eine geeignete Testumgebung für die Experimente aufgebaut werden?
- Wie werden relevante Testdaten der EDC Anwendung extrahiert und anonymisiert?
- Welche Python Bibliotheken für Graph ML Aufgaben gibt es?
- Welche Kategorien (Taxonomie) von Link Prediction Techniken gibt es?
- Welche Link Prediction Techniken werden in der Fallstudie untersucht, welche Python Bibliotheken kommen zum Einsatz?
- Wie werden die Ergebnisse der Untersuchungen bewertet und miteinander verglichen?
- Welche der untersuchten Varianten eignen sich für die Verknüpfung der Personendaten?

Methodik

- Für die explorative Arbeit wird eine agile Vorgehensweise nach SCRUM gewählt.
- Die **Anforderungen/Fragestellungen** werden in Form von **Stories** erfasst, mit dem Themensponsor abgeglichen und iterativ umgesetzt.
- Sobald eine Story umgesetzt ist, werden die **Resultate den Betreuern und dem Themensponsor präsentiert** und die nächsten Schritte geplant.
- Die Durchführung der Experimente orientiert sich am **Experimentierzyklus** aus dem Lehrplan 21 (Klett & Balmer, 2019).

Experimentierzyklus aus dem Lehrplan 21 (Klett & Balmer, 2019).

Grobplanung

2021/22	Themenantrag	Aufwand
Dez. 21	Themenwahl, erste Recherchen, WAW	2 Tage
Januar	Erste Einarbeitung, Antrag erstellen	2 Tage
Februar	Folienpräsentation Antrag und Präsentation	1 Tag
	Total	**5 Tage**

2022	Arbeit	Aufwand
März	Systematische Recherche, Literaturauswahl	2 Tage
April	Grundlagen Graphen und Link Prediction	10 Tage
Mai	Extraktion und Anonimisierung Testdaten EDC, Bereitstellung Umgebung, Auswahl Python Libraries	9 Tage
Juni	Taxonomie und erste Experimente Link Prediction	8 Tage
Juli	Bewertung und Vergleich Experimente	8 Tage
August	Bericht	4 Tage
September	Bericht und Bookbeitrag Abgabetermin: 26.09.22	4 Tage
	Total	**45 Tage**

Resultate

1. Wissensaufbau und Dokumentation **der Grundlagen von Graph Machine Learning**, so dass abgeschätzt werden kann, für welche Geschäftsanwendungen die Technologie in Zukunft eingesetzt werden kann.
2. Fallstudie mit **Vergleich verschiedener Link Prediction Ansätze** zur Verknüpfung von Personendaten der EDC Anwendung.
3. **Experimente** mit Source Code (Python und Jupiter Notebooks) zum Nachvollziehen der einzelnen Methoden und Verfahren.
4. **Schlussbericht**

Fragen und Antworten

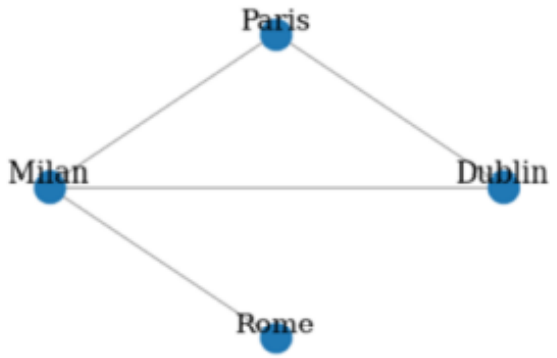
Herzlichen Dank für Ihre Aufmerksamkeit.

Bild: https://cutewallpaper.org

FAQ / Backup

Repräsentation von Graphen

Adjacency matrix



	Milan	Paris	Dublin	Rome
Milan	0	1	1	1
Paris	1	0	1	0
Dublin	1	1	0	0
Rome	1	0	0	0

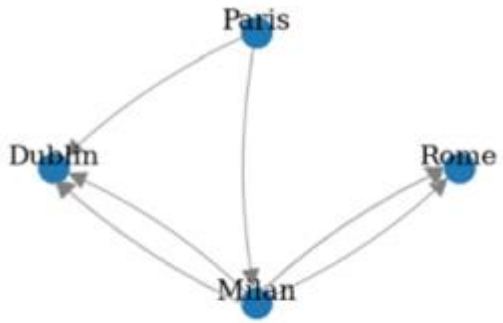
unidirected graph

	Milan	Paris	Dublin	Rome
Milan	0	0	1	1
Paris	1	0	1	0
Dublin	0	0	0	0
Rome	0	0	0	0

directed graph (digraph)

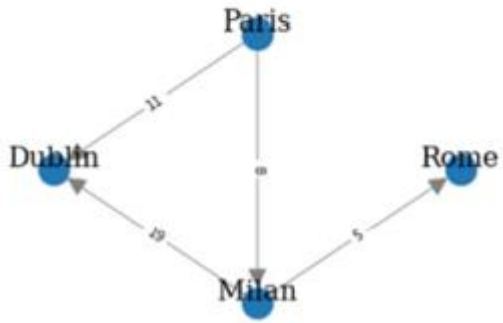
Quelle: Graph Machine Learing, ISBN 978-1-80020-449-2

Adjacency matrix II



	Milan	Paris	Dublin	Rome
Milan	0	0	2	2
Paris	1	0	1	0
Dublin	0	0	0	0
Rome	0	0	0	0

multigraph

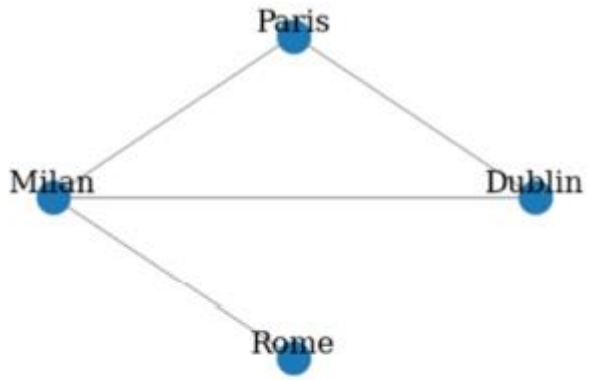


	Milan	Paris	Dublin	Rome
Milan	0	0	19	5
Paris	8	0	11	0
Dublin	0	0	0	0
Rome	0	0	0	0

weighted graph

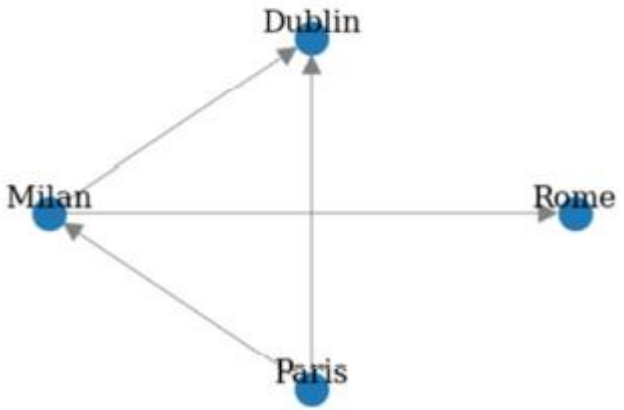
Quelle: Graph Machine Learing, ISBN 978-1-80020-449-2

Edge list



	Edge
1	{Milan, Dublin}
2	{Milan, Paris}
3	{Paris, Dublin}
4	{Milan, Rome}

unidirected graph

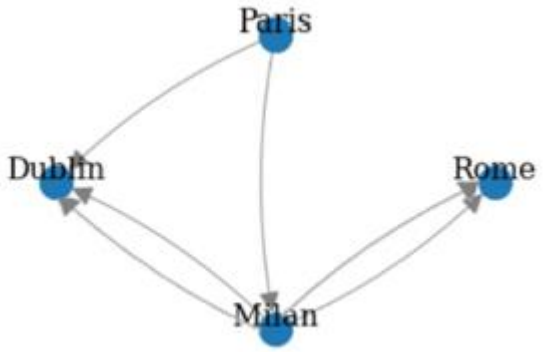


	Edge
1	(Milan, Dublin)
2	(Paris, Milan)
3	(Paris, Dublin)
4	(Milan, Rome)

directed graph (digraph)

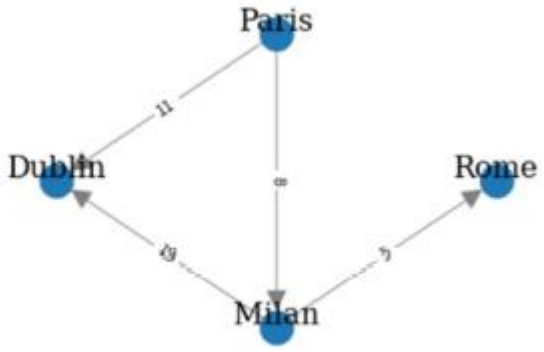
Quelle: Graph Machine Learing, ISBN 978-1-80020-449-2

Edge list II



	Edge
1	(Milan, Dublin)
2	(Milan, Dublin)
3	(Milan, Rome)
4	(Milan, Rome)
5	(Paris, Dublin)
6	(Paris, Milan)

multigraph



	Edge	Weight
1	(Milan, Dublin)	19
2	(Paris, Milan)	8
3	(Paris, Dublin)	11
4	(Milan, Rome)	5

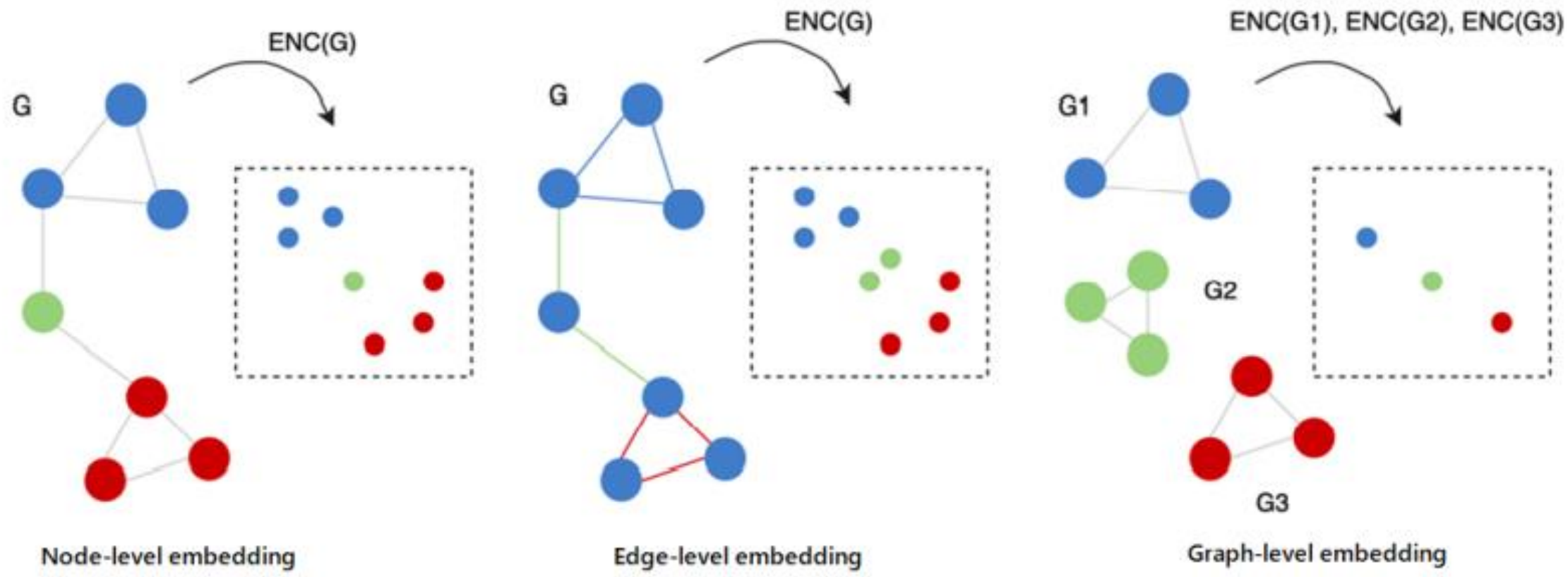
weighted graph

Quelle: Graph Machine Learing, ISBN 978-1-80020-449-2

Graph Machine Learning

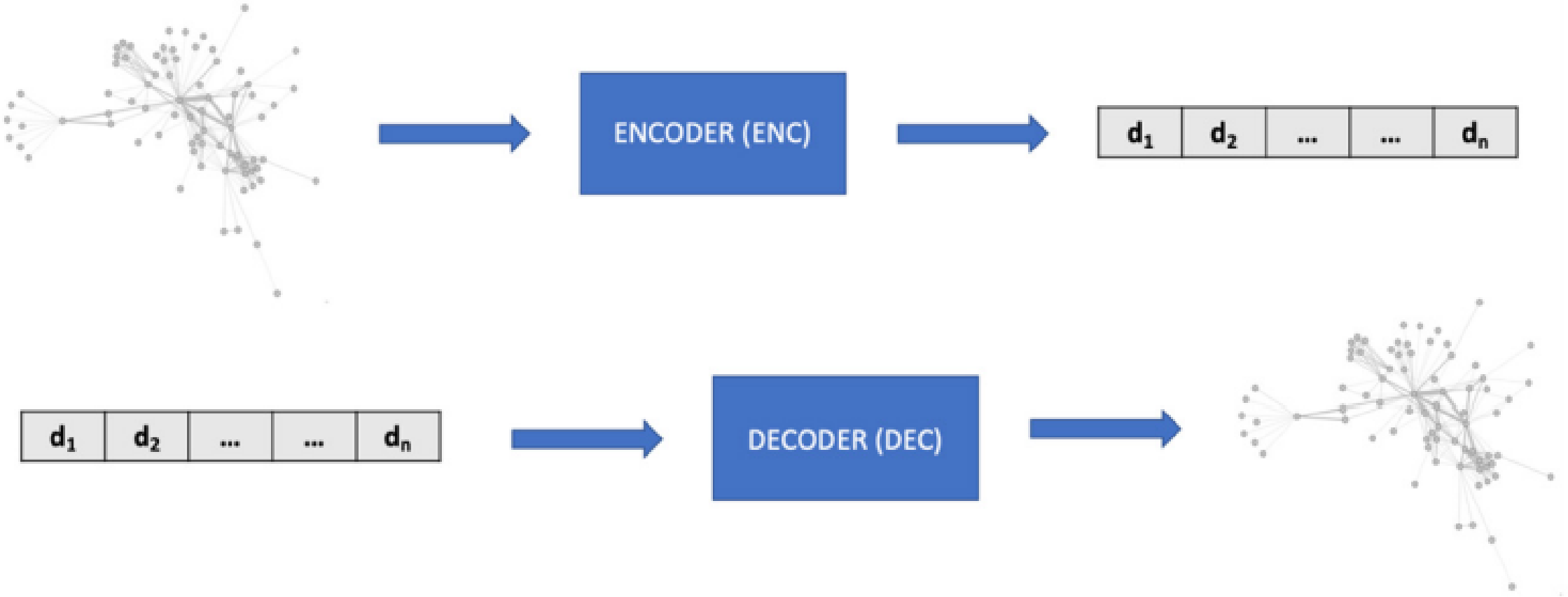
Berner Fachhochschule | Haute école spécialisée bernoise | Bern University of Applied Sciences

Embeddings



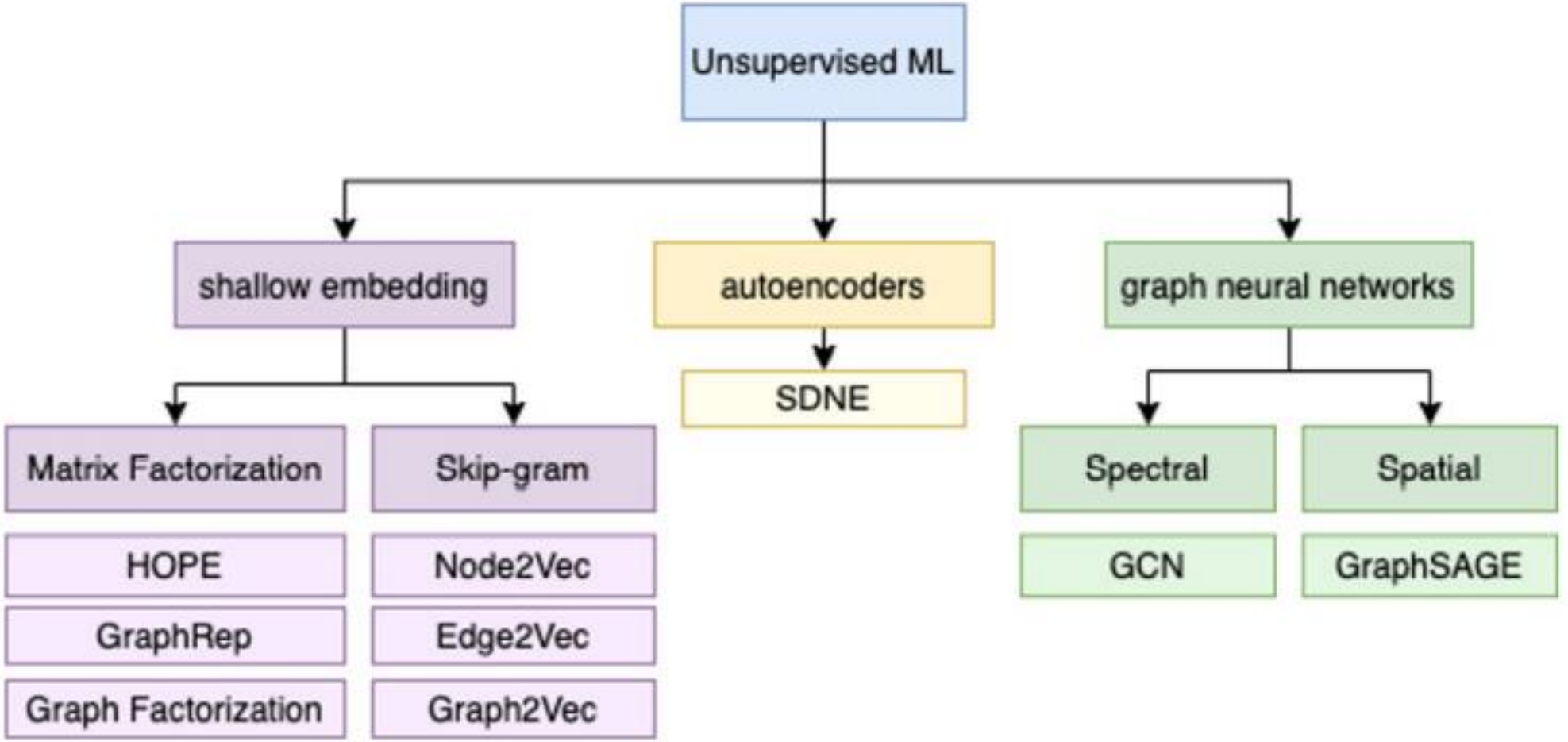
Quelle: Graph Machine Learing, ISBN 978-1-80020-449-2

Encoder / Decoder Architektur



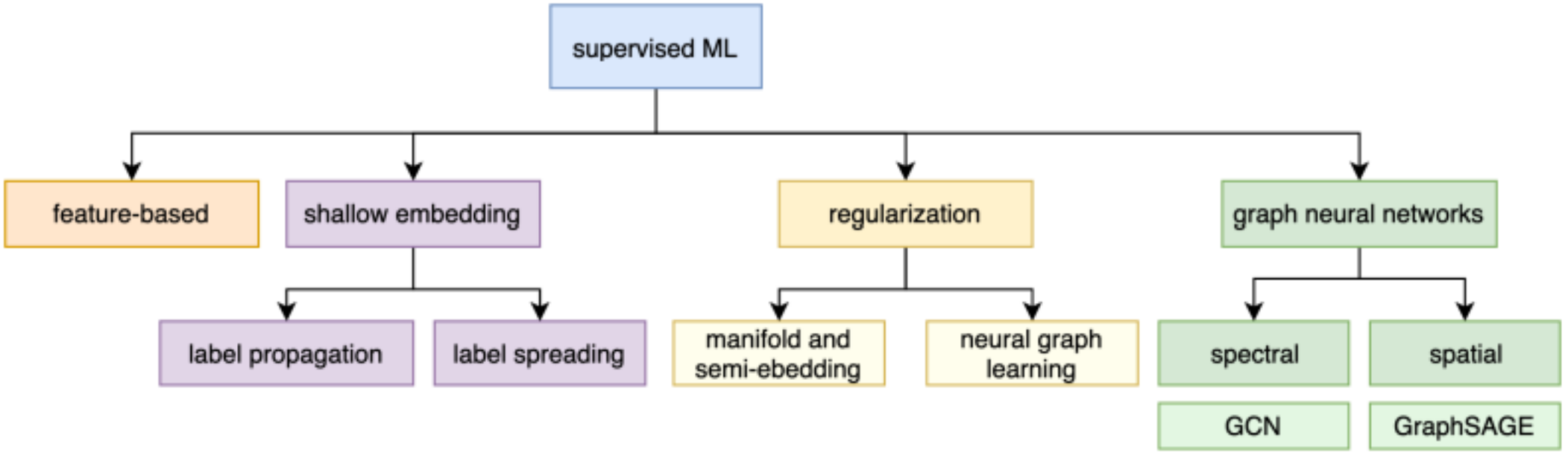
Quelle: Graph Machine Learing, ISBN 978-1-80020-449-2

Unsupervised ML



Quelle: Graph Machine Learing, ISBN 978-1-80020-449-2

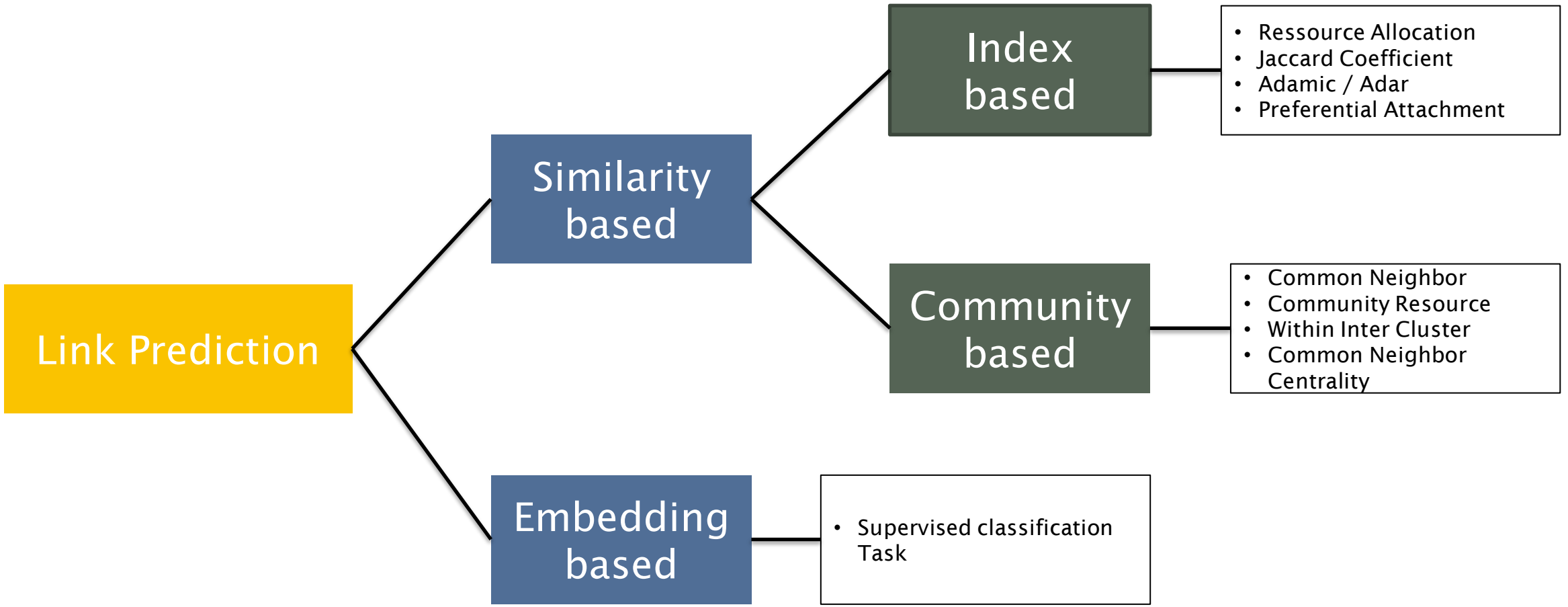
Supervised ML



Quelle: Graph Machine Learing, ISBN 978-1-80020-449-2

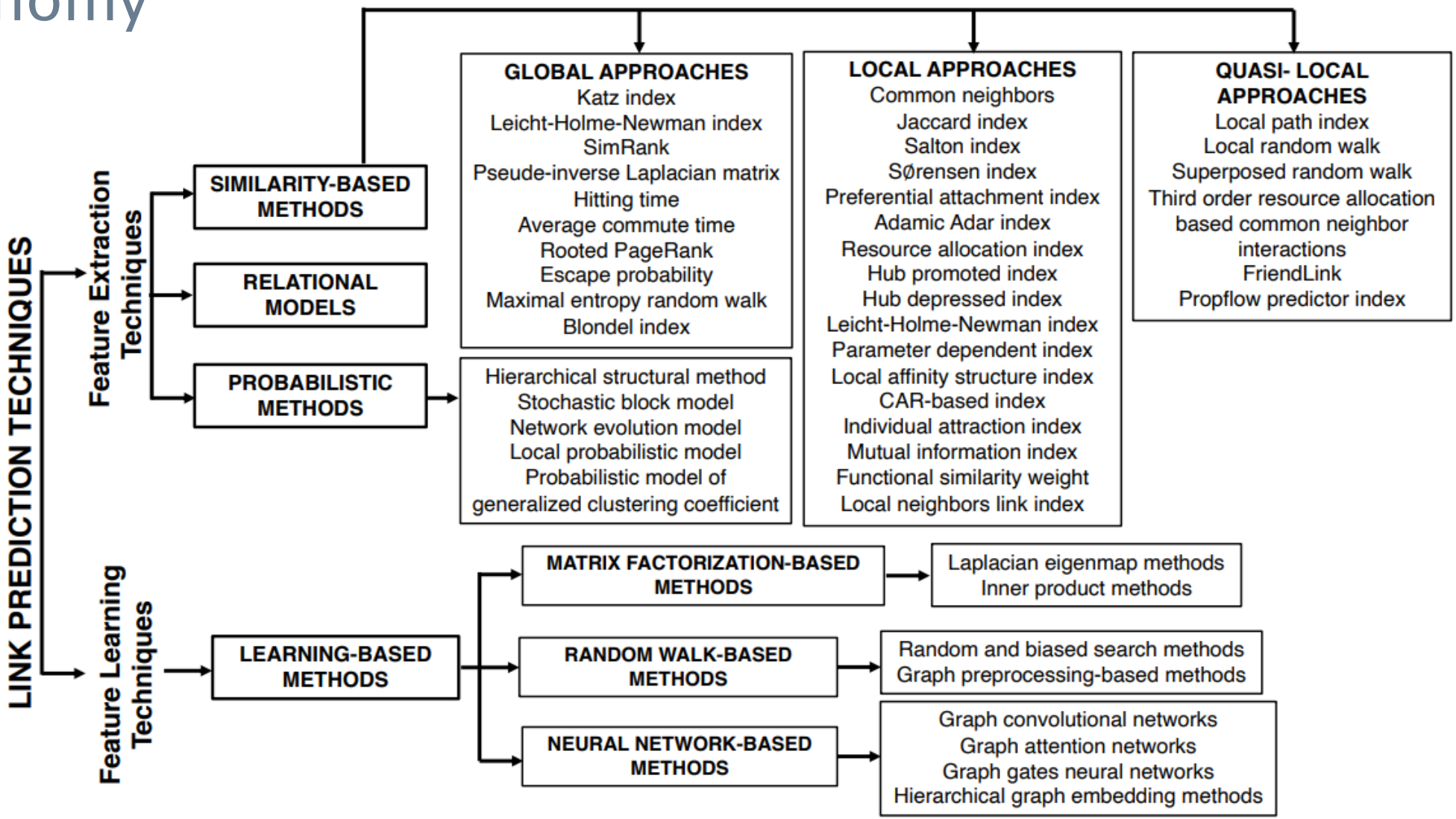
Link Prediction

Taxonomy



Quelle: Graph Machine Learing, ISBN 978-1-80020-449-2

Taxonomy



Quelle: https://arxiv.org/abs/1901.03425v5

The end.

Bild: https://cutewallpaper.org