LAMARR
Institute for Machine Learning
and Artificial Intelligence

universitätbonn

Fraunhofer

TU WIEN
TECHNISCHE
UNIVERSITÄT
WIEN

Universität Potsdam

Hidden Schema Networks

Ramsés J. Sánchez, Lukas Conrads, Pascal Welke, Kostadin Cvejoski and César Ojeda

Large Language Models infer representations that **implicitly encode** rich contextual word **semantics** and sentence-level **grammar**

Large Language Models infer representations that **implicitly encode** rich contextual word **semantics** and sentence-level **grammar**

A Structural Probe for Finding Syntax in Word Representations
John Hewitt Christopher D. Manning (2019)

Large Language Models infer representations that **implicitly encode** rich contextual word **semantics** and sentence-level **grammar**

Open Sesame: Getting Inside BERT's Linguistic Knowledge

Yongjie Lin[a,*] and Yi Chern Tan[a,*] and Robert Frank[b]

(2019)

A Structural Probe for Finding Syntax in Word Representations

John Hewitt Christopher D. Manning (2019)

WHAT DO YOU LEARN FROM CONTEXT? PROBING FOR SENTENCE STRUCTURE IN CONTEXTUALIZED WORD REPRESENTATIONS

Ian Tenney,[*1] Patrick Xia,[2] Berlin Chen,[3] Alex Wang,[4] Adam Poliak,[2]
R. Thomas McCoy,[2] Najoung Kim,[2] Benjamin Van Durme,[2] Samuel R. Bowman,[4]
Dipanjan Das,[1] and Ellie Pavlick[1,5]

(2019)

Large Language Models infer representations that **implicitly encode** rich contextual word **semantics** and sentence-level **grammar**

Open Sesame: Getting Inside BERT's Linguistic Knowledge

Yongjie Lin[a,*] and Yi Chern Tan[a,*] and Robert Frank[b]

(2019)

A Structural Probe for Finding Syntax in Word Representations

John Hewitt Christopher D. Manning (2019)

WHAT DO YOU LEARN FROM CONTEXT? PROBING FOR SENTENCE STRUCTURE IN CONTEXTUALIZED WORD REPRESENTATIONS

Ian Tenney,[*1] Patrick Xia,[2] Berlin Chen,[3] Alex Wang,[4] Adam Poliak,[2]
R. Thomas McCoy,[2] Najoung Kim,[2] Benjamin Van Durme,[2] Samuel R. Bowman,[4]
Dipanjan Das,[1] and Ellie Pavlick[1,5]

(2019)

How Can We Know What Language Models Know?

Zhengbao Jiang[1*] Frank F. Xu[1*] Jun Araki[2] Graham Neubig[1] (2020)

Large Language Models infer representations that **implicitly encode** rich contextual word **semantics** and sentence-level **grammar**

Open Sesame: Getting Inside BERT's Linguistic Knowledge

Yongjie Lin[a,*] and Yi Chern Tan[a,*] and Robert Frank[b]

(2019)

A Structural Probe for Finding Syntax in Word Representations

John Hewitt Christopher D. Manning (2019)

WHAT DO YOU LEARN FROM CONTEXT? PROBING FOR SENTENCE STRUCTURE IN CONTEXTUALIZED WORD REPRESENTATIONS

Ian Tenney,[*1] Patrick Xia,[2] Berlin Chen,[3] Alex Wang,[4] Adam Poliak,[2] R. Thomas McCoy,[2] Najoung Kim,[2] Benjamin Van Durme,[2] Samuel R. Bowman,[4] Dipanjan Das,[1] and Ellie Pavlick[1,5]

(2019)

How Can We Know What Language Models Know?

Zhengbao Jiang[1*] Frank F. Xu[1*] Jun Araki[2] Graham Neubig[1] (2020)

Large Language Models infer representations that **implicitly encode** rich contextual word **semantics** and sentence-level **grammar**

Open Sesame: Getting Inside BERT's Linguistic Knowledge

Yongjie Lin[a,*] and Yi Chern Tan[a,*] and Robert Frank[b]

(2019)

Syntactic Structure from Deep Learning

Tal Linzen[1] and Marco Baroni[2,3,4] (2021)

A Structural Probe for Finding Syntax in Word Representations

John Hewitt Christopher D. Manning (2019)

WHAT DO YOU LEARN FROM CONTEXT? PROBING FOR SENTENCE STRUCTURE IN CONTEXTUALIZED WORD REPRESENTATIONS

Ian Tenney,[*1] Patrick Xia,[2] Berlin Chen,[3] Alex Wang,[4] Adam Poliak,[2] R. Thomas McCoy,[2] Najoung Kim,[2] Benjamin Van Durme,[2] Samuel R. Bowman,[4] Dipanjan Das,[1] and Ellie Pavlick[1,5]

(2019)

How Can We Know What Language Models Know?

Zhengbao Jiang[1*] Frank F. Xu[1*] Jun Araki[2] Graham Neubig[1] (2020)

What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models

Allyson Ettinger (2020)

Large Language Models infer representations that **implicitly encode** rich contextual word **semantics** and sentence-level **grammar**

Open Sesame: Getting Inside BERT's Linguistic Knowledge

Yongjie Lin[a,*] and Yi Chern Tan[a,*] and Robert Frank[b]

(2019)

A Structural Probe for Finding Syntax in Word Representations

John Hewitt Christopher D. Manning (2019)

Syntactic Structure from Deep Learning

Tal Linzen[1] and Marco Baroni[2,3,4] (2021)

WHAT DO YOU LEARN FROM CONTEXT? PROBING FOR SENTENCE STRUCTURE IN CONTEXTUALIZED WORD REPRESENTATIONS

Ian Tenney,[*1] Patrick Xia,[2] Berlin Chen,[3] Alex Wang,[4] Adam Poliak,[2] R. Thomas McCoy,[2] Najoung Kim,[2] Benjamin Van Durme,[2] Samuel R. Bowman,[4] Dipanjan Das,[1] and Ellie Pavlick[1,5]

(2019)

How Can We Know What Language Models Know?

Zhengbao Jiang[1*] Frank F. Xu[1*] Jun Araki[2] Graham Neubig[1] (2020)

What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models

Allyson Ettinger (2020)

Large Language Models infer representations that **implicitly encode** rich contextual word **semantics** and sentence-level **grammar**

Open Sesame: Getting Inside BERT's Linguistic Knowledge

Yongjie Lin[a,*] and Yi Chern Tan[a,*] and Robert Frank[b]

(2019)

A Structural Probe for Finding Syntax in Word Representations

John Hewitt Christopher D. Manning (2019)

Syntactic Structure from Deep Learning

Tal Linzen[1] and Marco Baroni[2,3,4] (2021)

Emergent linguistic structure in artificial neural networks trained by self-supervision

Christopher D. Manning[a,1], Kevin Clark[a], John Hewitt[a], Urvashi Khandelwal[a], and Omer Levy[b] (2020)

Large Language Models **struggle** to solve tasks that require **formal** and **commonsense reasoning**

Are NLP Models really able to Solve Simple Math Word Problems?

Arkil Patel Satwik Bhattamishra Navin Goyal

(2021)

Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly

Nora Kassner, Hinrich Schütze

(2020)

LARGE LANGUAGE MODELS ARE NOT ZERO-SHOT COMMUNICATORS

(2022)

Laura Ruis[1], Akbir Khan[1], Stella Biderman[2,3], Sara Hooker[4], Tim Rocktäschel[1], Edward Grefenstette[1,5]

Large Language Models **struggle** to solve tasks that require **formal** and **commonsense reasoning**

Things not Written in Text: Exploring Spatial Commonsense from Visual Signals

Xiao Liu[1], Da Yin[2], Yansong Feng[1,3*] and Dongyan Zhao[1,4,5] (2022)

On the Paradox of Learning to Reason from Data

Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, Guy Van den Broeck (2022)

COMPS: Conceptual Minimal Pair Sentences for testing Robust Property Knowledge and its Inheritance in Pre-trained Language Models

Kanishka Misra Julia Rayz Allyson Ettinger

(2023)

Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change)

Karthik Valmeekam* (2023) Alberto Olmo*

Sarath Sreedharan† Subbarao Kambhampati

Large Language Models can be guided to generate reasoning explicitly: **Chain-of-Thought**

Rethinking with Retrieval: Faithful Large Language Model Inference

Hangfeng He[†*] Hongming Zhang[‡] Dan Roth[§]

(2022)

Iteratively Prompt Pre-trained Language Models for Chain of Thought

Boshi Wang, Xiang Deng and Huan Sun

(2022)

Large Language Models can be guided to generate reasoning explicitly: **Chain-of-Thought**

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou

(2022)

Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering

Pan Lu[1,3], Swaroop Mishra[2,3], Tony Xia[1], Liang Qiu[1], Kai-Wei Chang[1], Song-Chun Zhu[1], Oyvind Tafjord[3], Peter Clark[3], Ashwin Kalyan[3]

(2022)

Improving mathematical reasoning with process supervision

(2023) OpenAI

We propose to use Large Language Models **to infer** unsupervised **representations for reasoning**

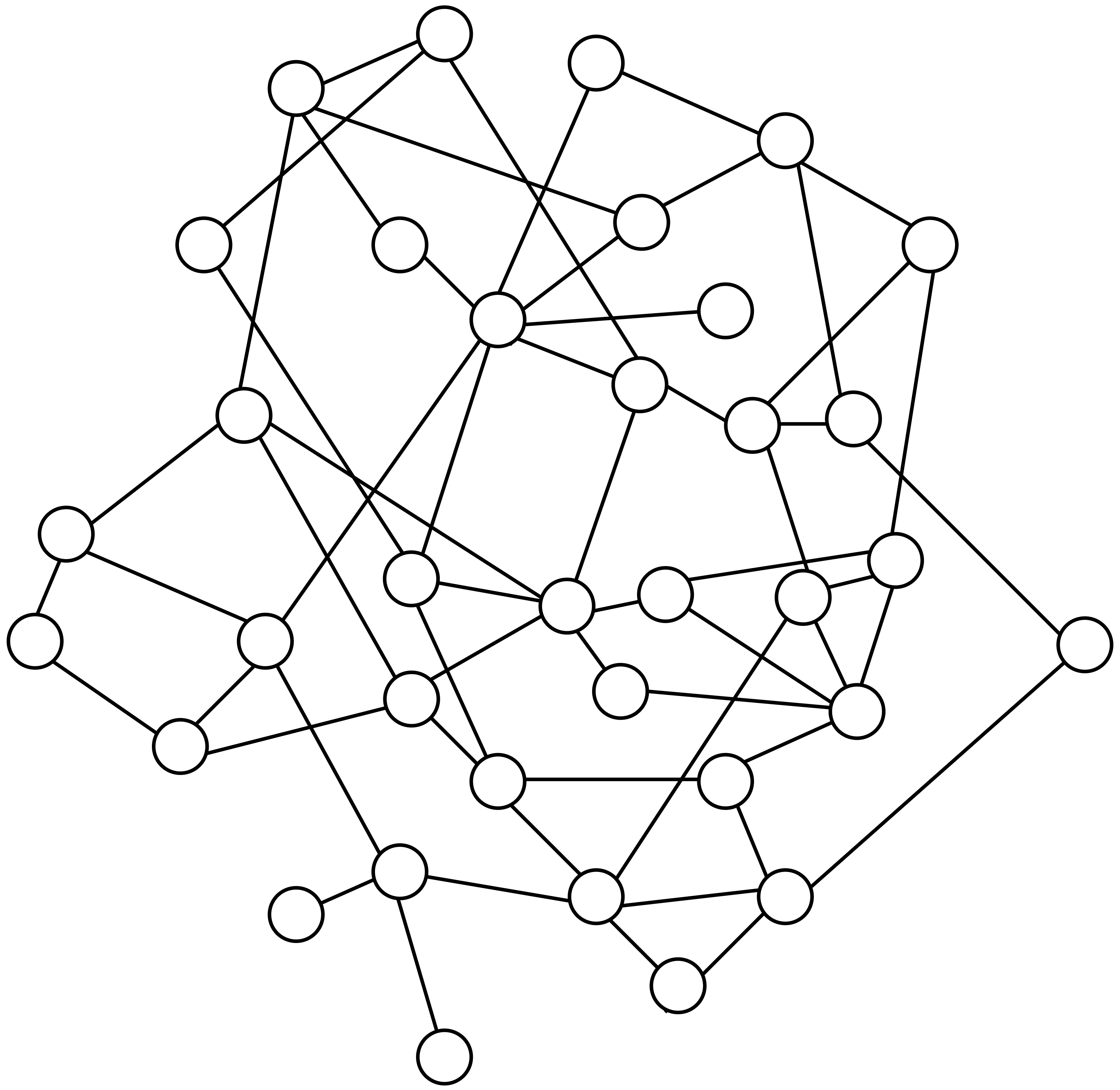
We propose to use Large Language Models **to infer** unsupervised **representations for reasoning**

Minimal inductive biases → **Relational structures** that allow for **compositionality**

Hidden Schema Networks

We assume

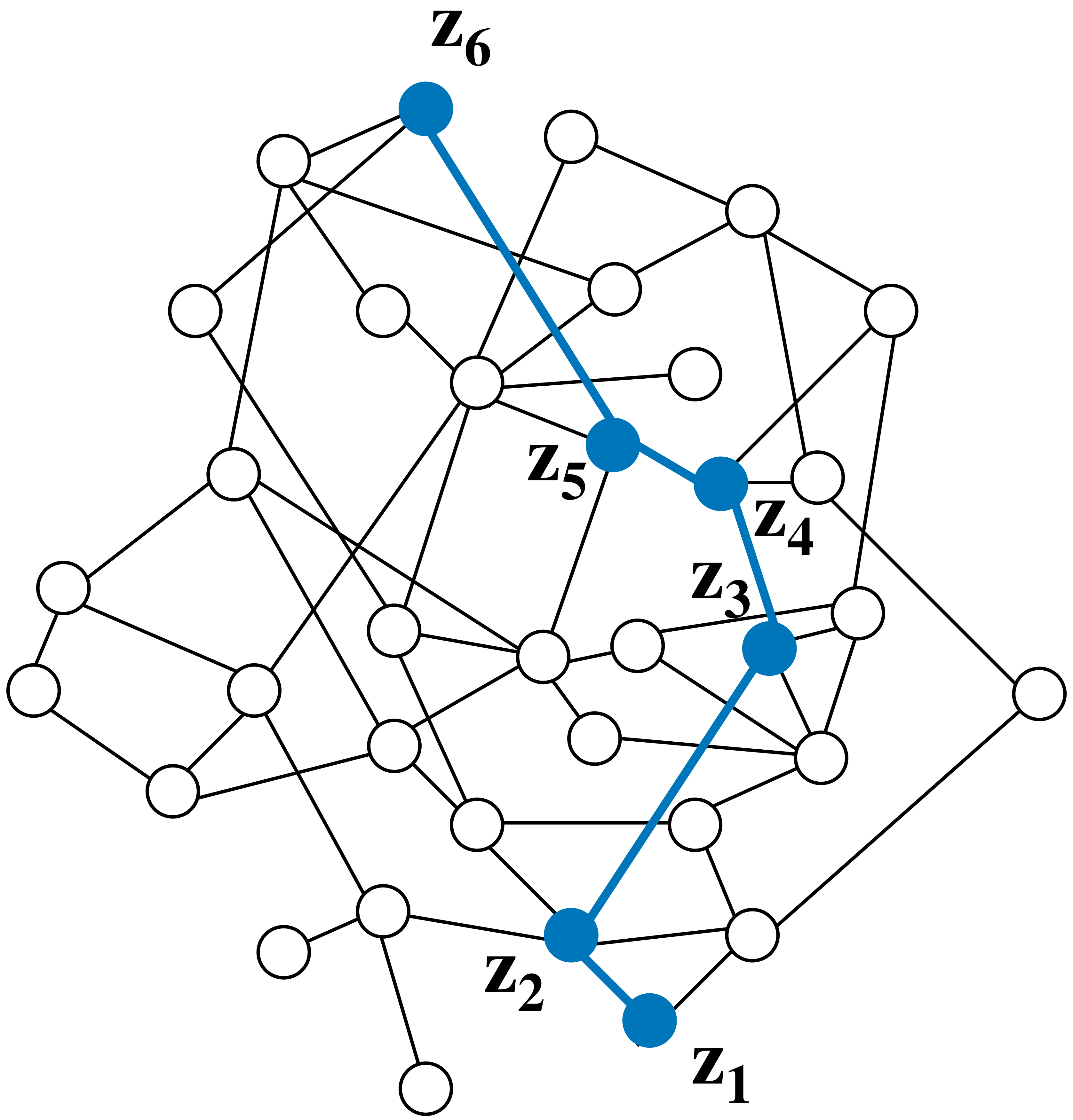
1. There is a set of symbols encoding some high-level, abstract semantic content of natural language



Hidden Schema Networks

We assume

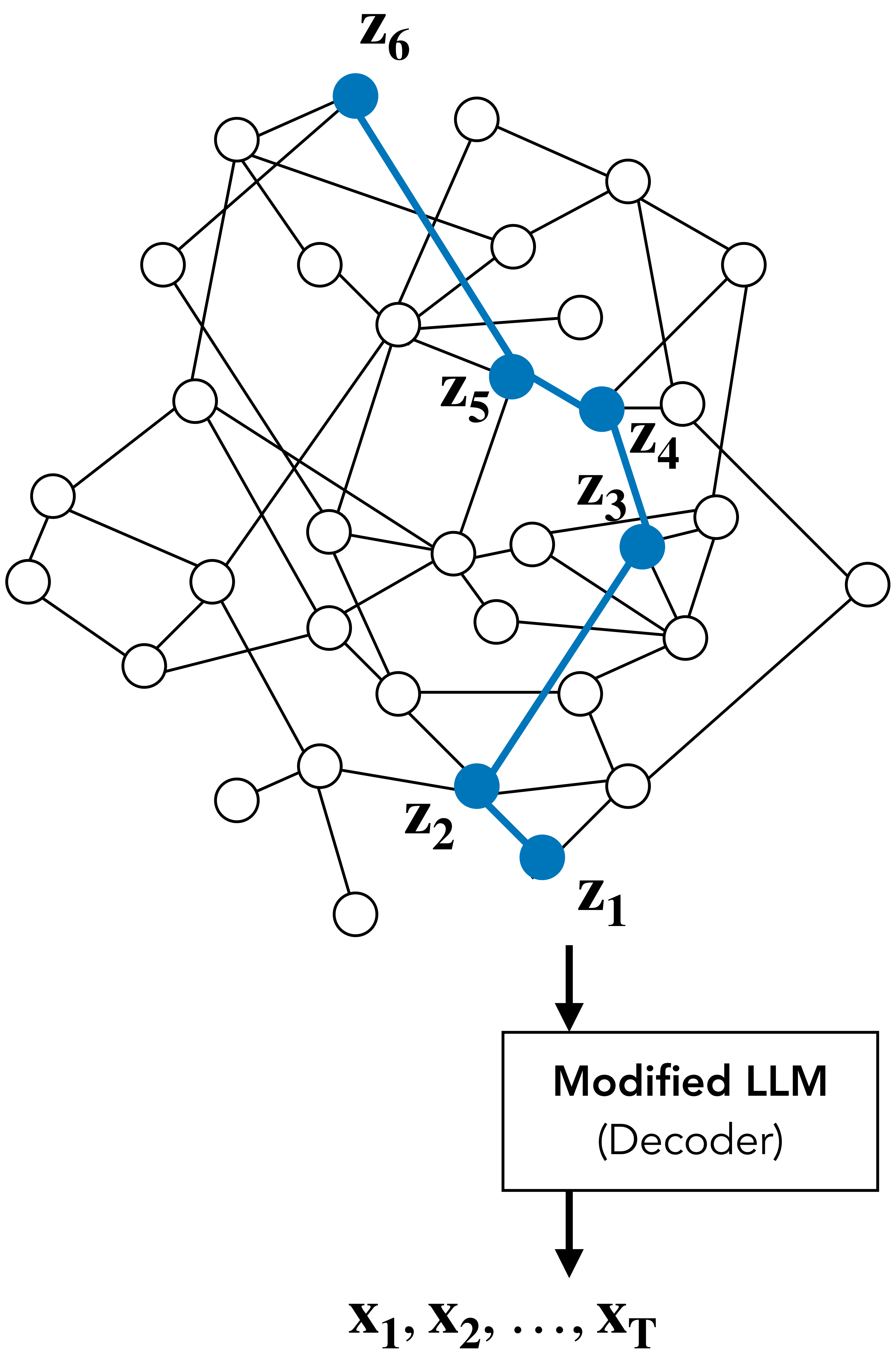
1. There is a set of symbols encoding some high-level, abstract semantic content of natural language
2. The **schemata** are sequences of connected symbols, composed by random walkers



Hidden Schema Networks

We assume

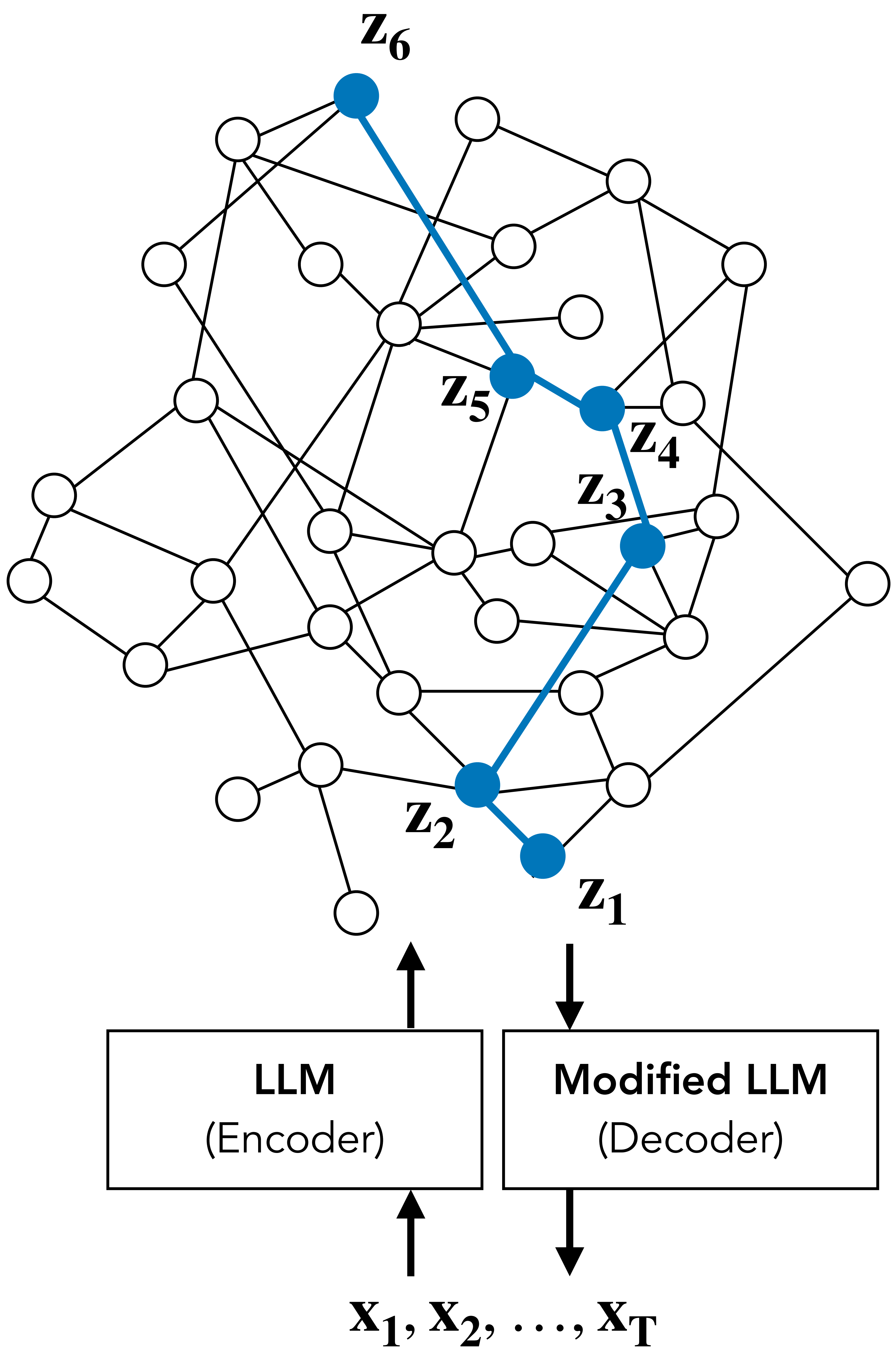
1. There is a set of symbols encoding some high-level, abstract semantic content of natural language
2. The **schemata** are sequences of connected symbols, composed by random walkers
3. Sentences are generated conditioned on the schemata



Hidden Schema Networks

We assume

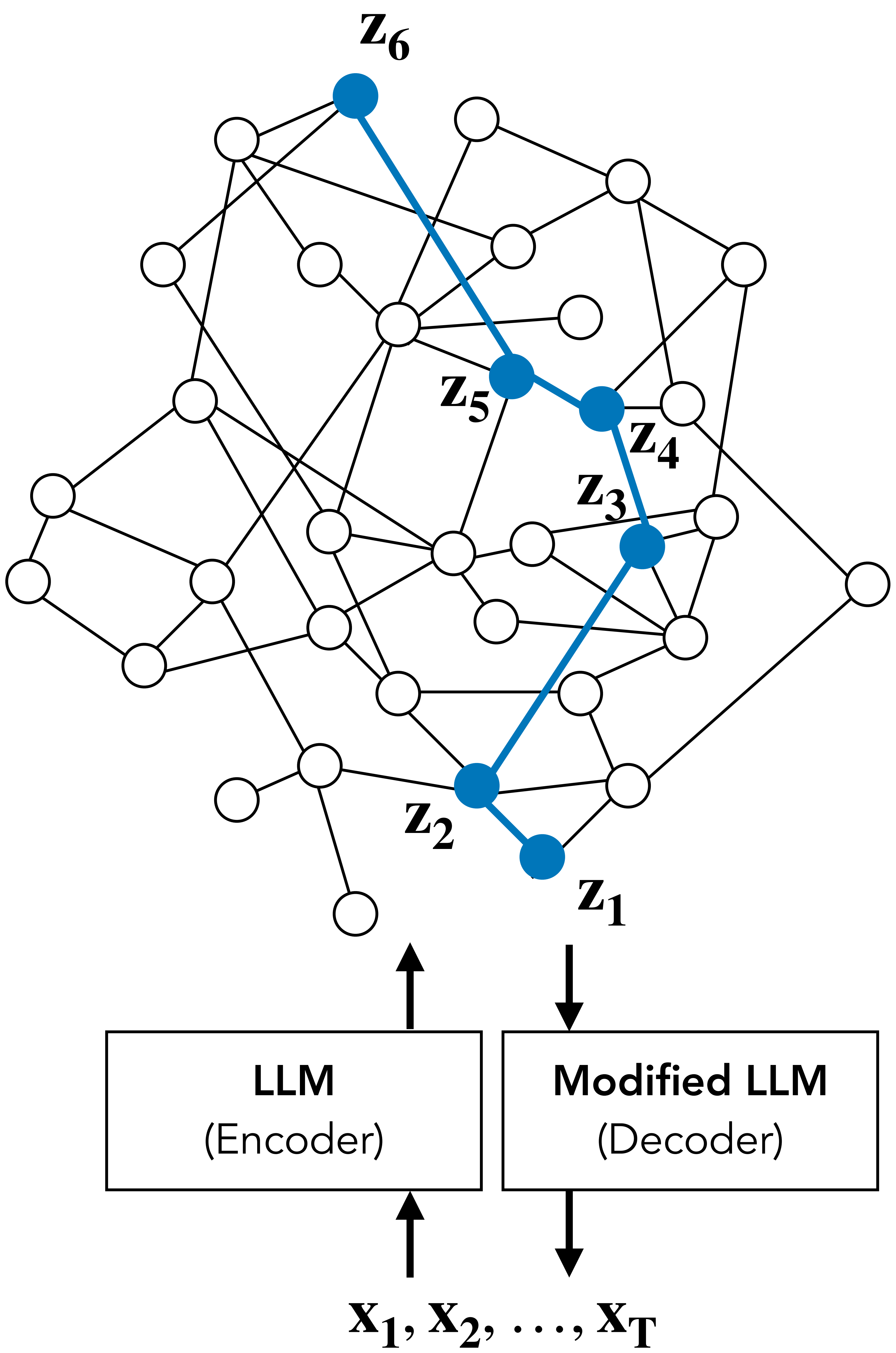
1. There is a set of symbols encoding some high-level, abstract semantic content of natural language
2. The **schemata** are sequences of connected symbols, composed by random walkers
3. Sentences are generated conditioned on the schemata



Hidden Schema Networks

We assume

1. There is a set of symbols encoding some high-level, abstract semantic content of natural language
2. The **schemata** are sequences of connected symbols, composed by random walkers
3. Sentences are generated conditioned on the schemata



We infer

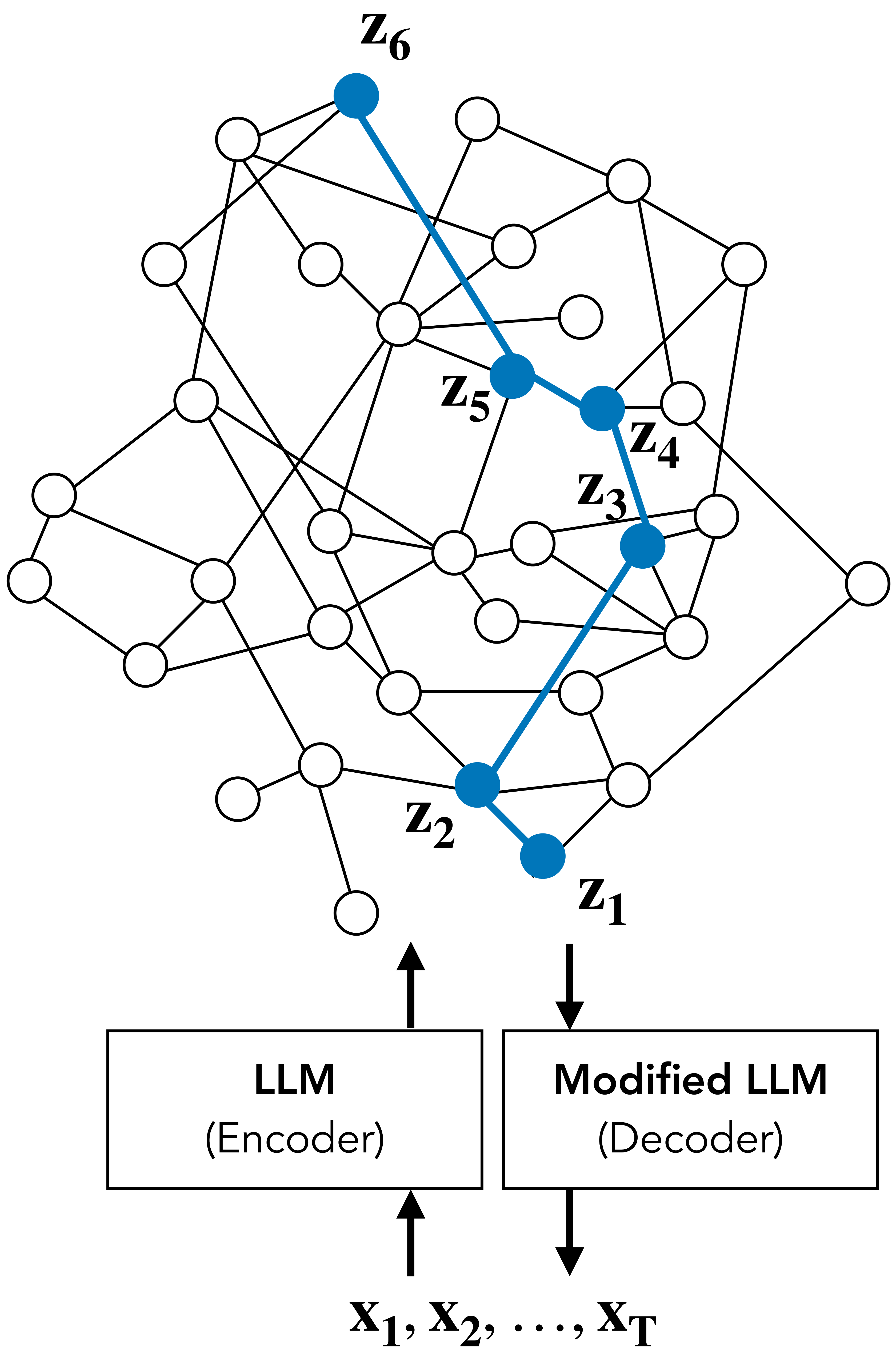
1. $q_\phi(\mathbf{A})$

   Posterior distribution over **global** graph

Hidden Schema Networks

We assume

1. There is a set of symbols encoding some high-level, abstract semantic content of natural language
2. The **schemata** are sequences of connected symbols, composed by random walkers
3. Sentences are generated conditioned on the schemata



We infer

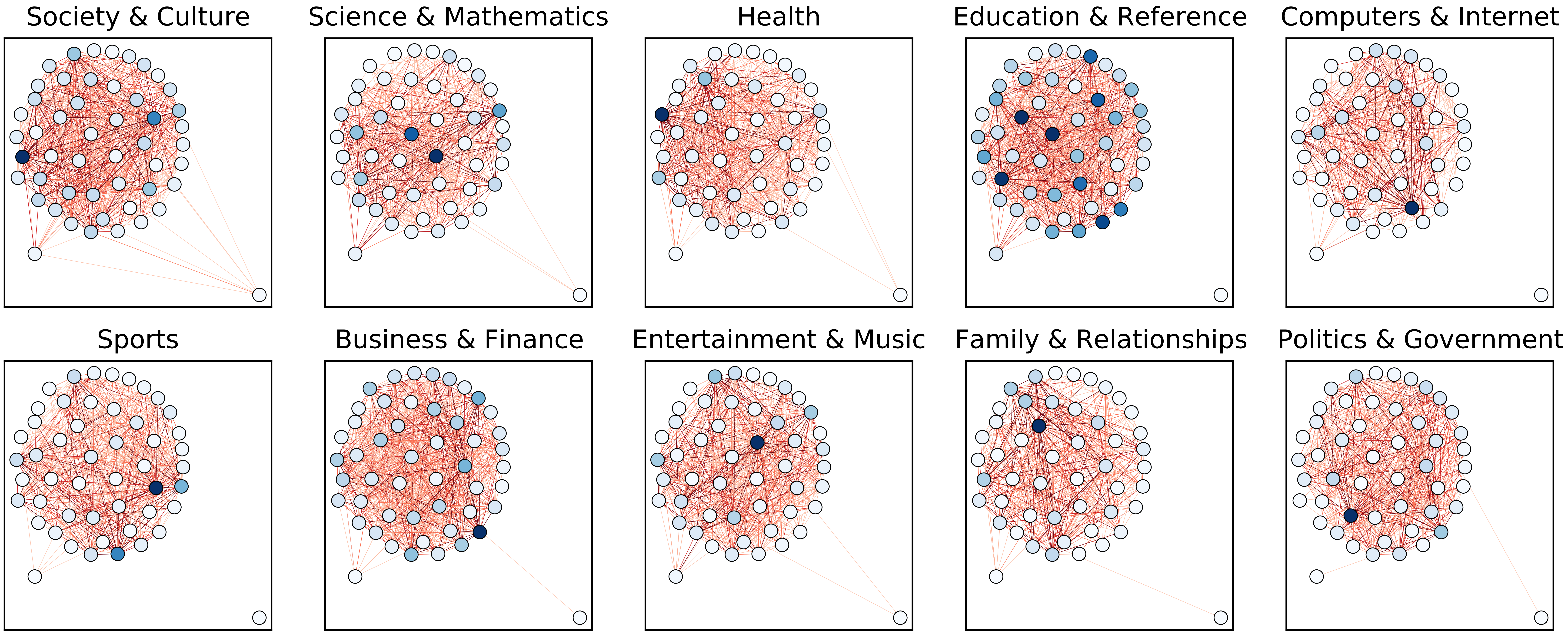
1. $q_\phi(\mathbf{A})$

   Posterior distribution over **global** graph

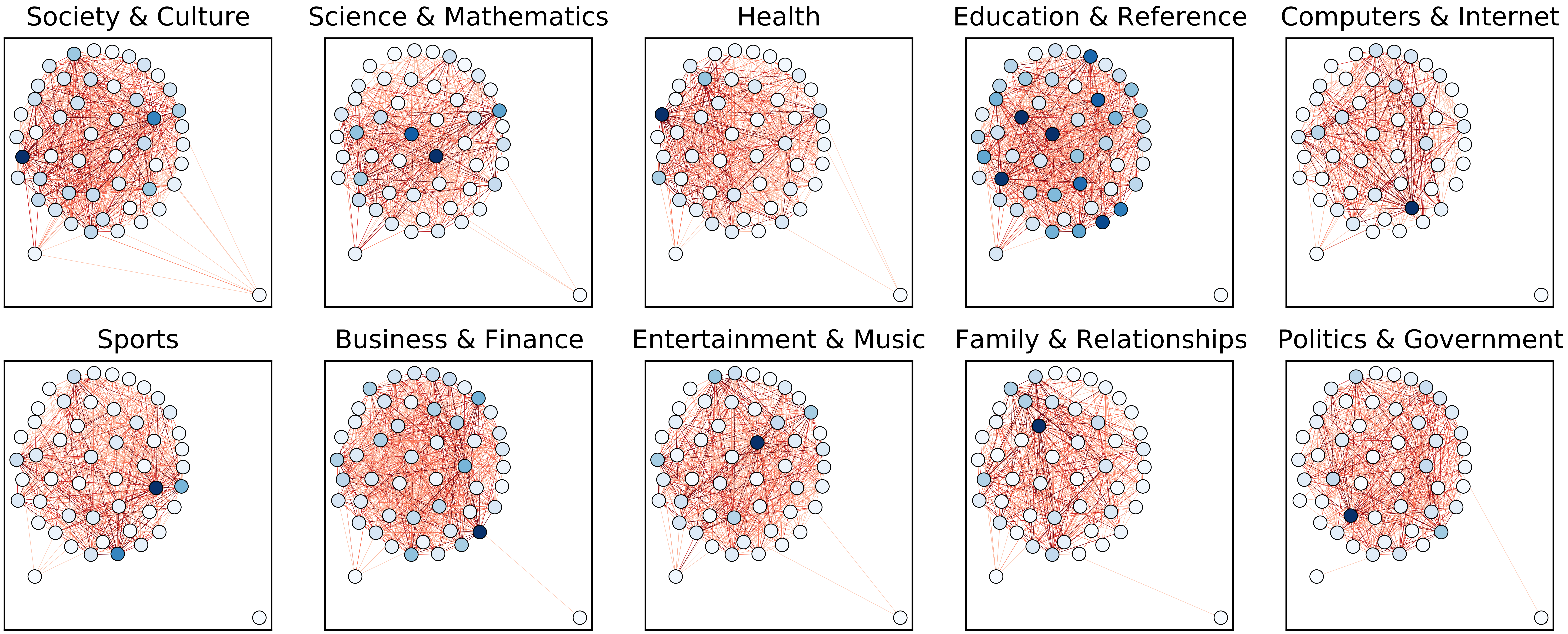
2. $q_\phi(\mathbf{z}_{1:L} | \mathbf{x}_{1:T}, \mathbf{A})$

   Posterior distribution over **local** random walks (schemata)

Hidden Schema Networks inferred from Yahoo



Hidden Schema Networks inferred from Yahoo



Symbols are interpreted a **posteriori**

Hidden Schema Networks for commonsense reasoning

Subject	Relation	Object
PersonX makes PersonY's coffee	xIntent	PersonX wanted to be helpful

TASK: given **Subject + Relation** generate **Object**

Hidden Schema Networks for commonsense reasoning

Subject	Relation	Object
PersonX makes PersonY's coffee	xIntent	PersonX wanted to be helpful

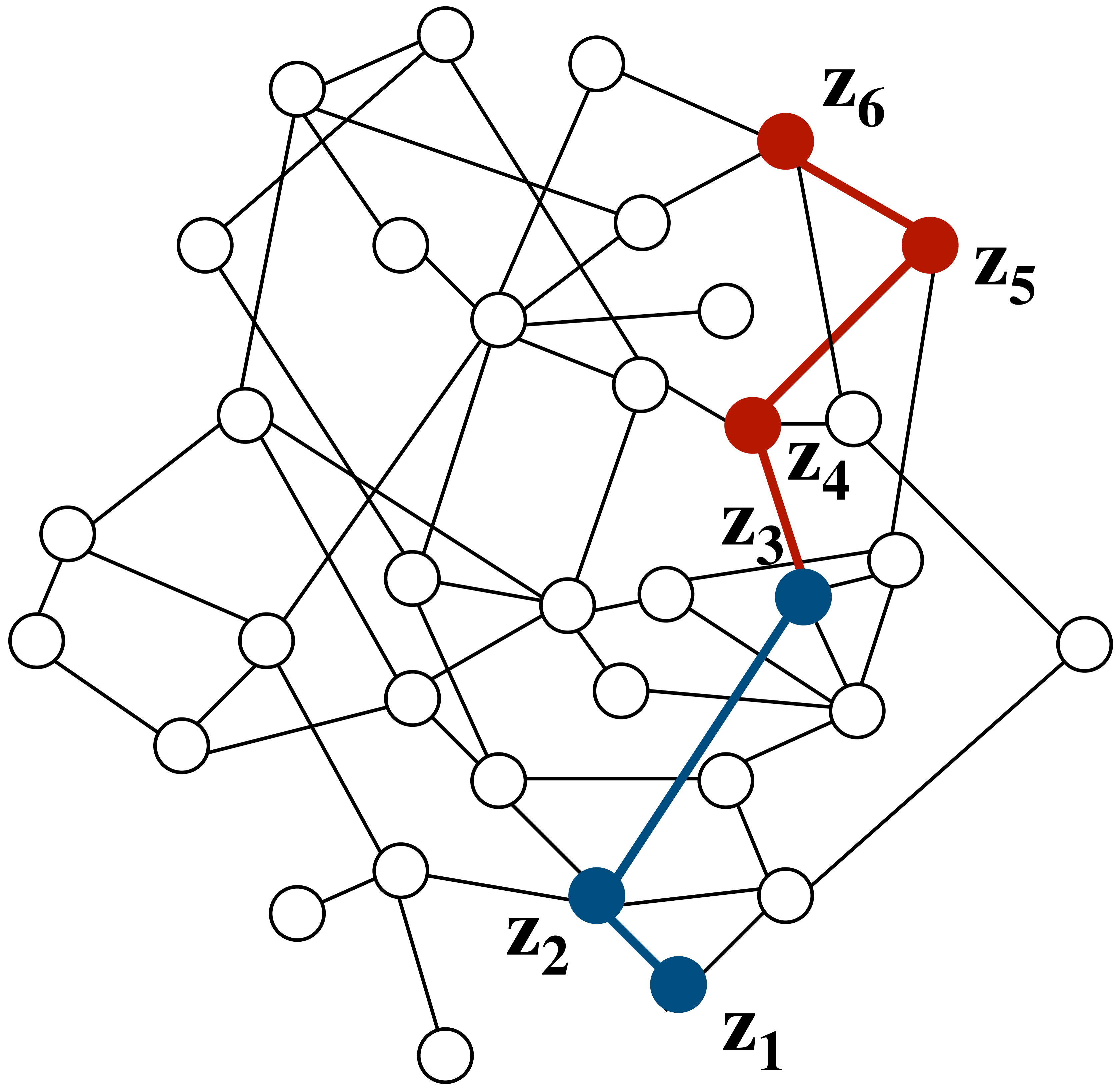
TASK: given **Subject + Relation** generate **Object**

Reasoning with **Hidden Schemata**

1. Encode **s + r + o** onto random walks

(s, r, o)

(s, r)



Hidden Schema Networks for commonsense reasoning

Subject	Relation	Object
PersonX makes PersonY's coffee	xIntent	PersonX wanted to be helpful

TASK: given **Subject + Relation** generate **Object**

Reasoning with **Hidden Schemata**

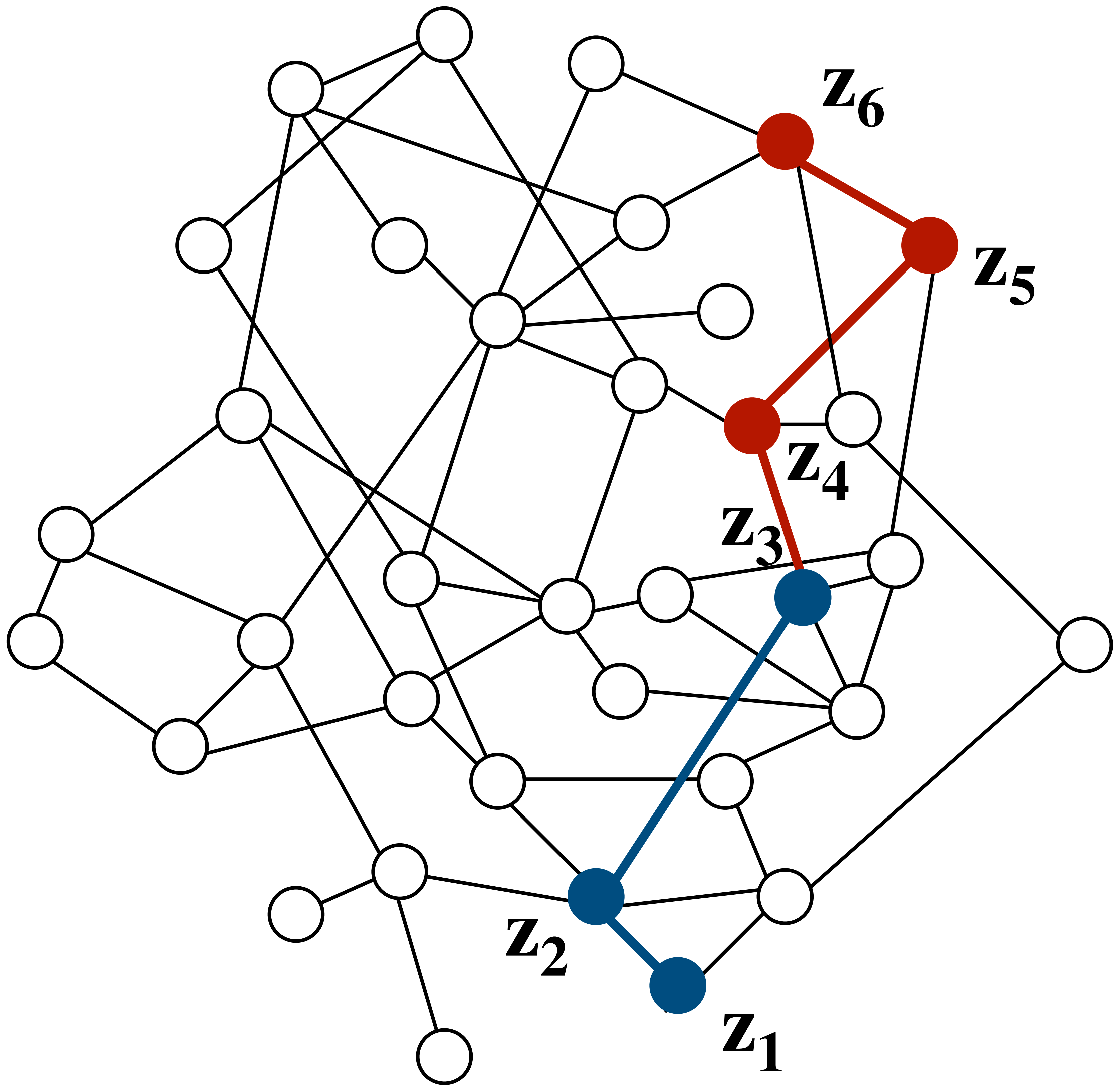
1. Encode **s + r + o** onto random walks

(s, r, o)

(s, r)

2. Train ``reasoning" autoregressive models on **2nd half of random walks** (the half encoding **o**)

reasoning



Hidden Schema Networks for commonsense reasoning

Subject	Relation	Object
PersonX makes PersonY's coffee	xIntent	PersonX wanted to be helpful

TASK: given **Subject + Relation** generate **Object**

Reasoning with **Hidden Schemata**

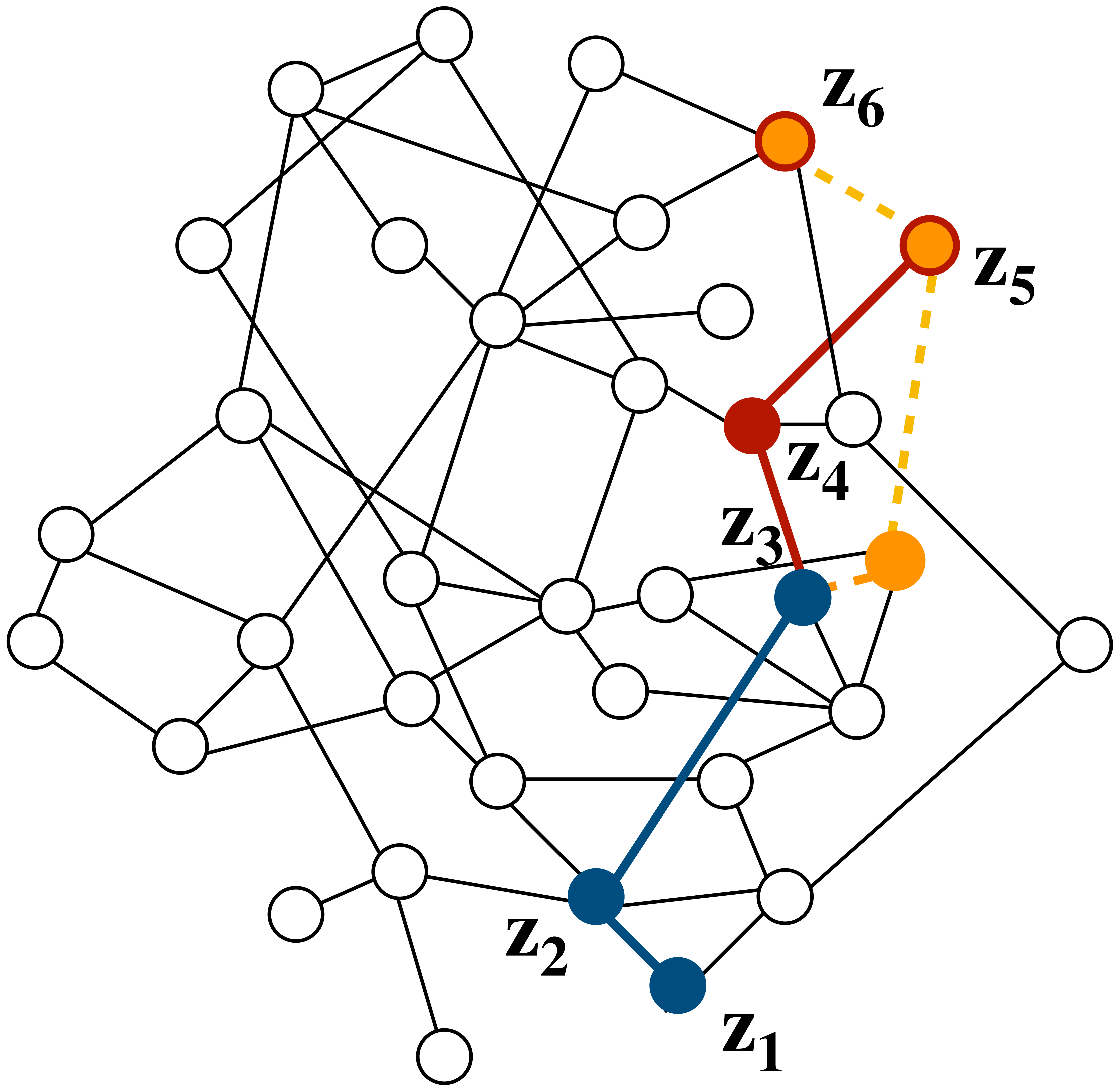
1. Encode **s + r + o** onto random walks

(s, r, o)

(s, r)

2. Train ``reasoning" autoregressive models on **2nd half of random walks** (the half encoding **o**)

reasoning



Hidden Schema Networks for commonsense reasoning

Subject	Relation	Object
PersonX makes PersonY's coffee	xIntent	PersonX wanted to be helpful

TASK: given **Subject + Relation** generate **Object**

Reasoning with **Hidden Schemata**

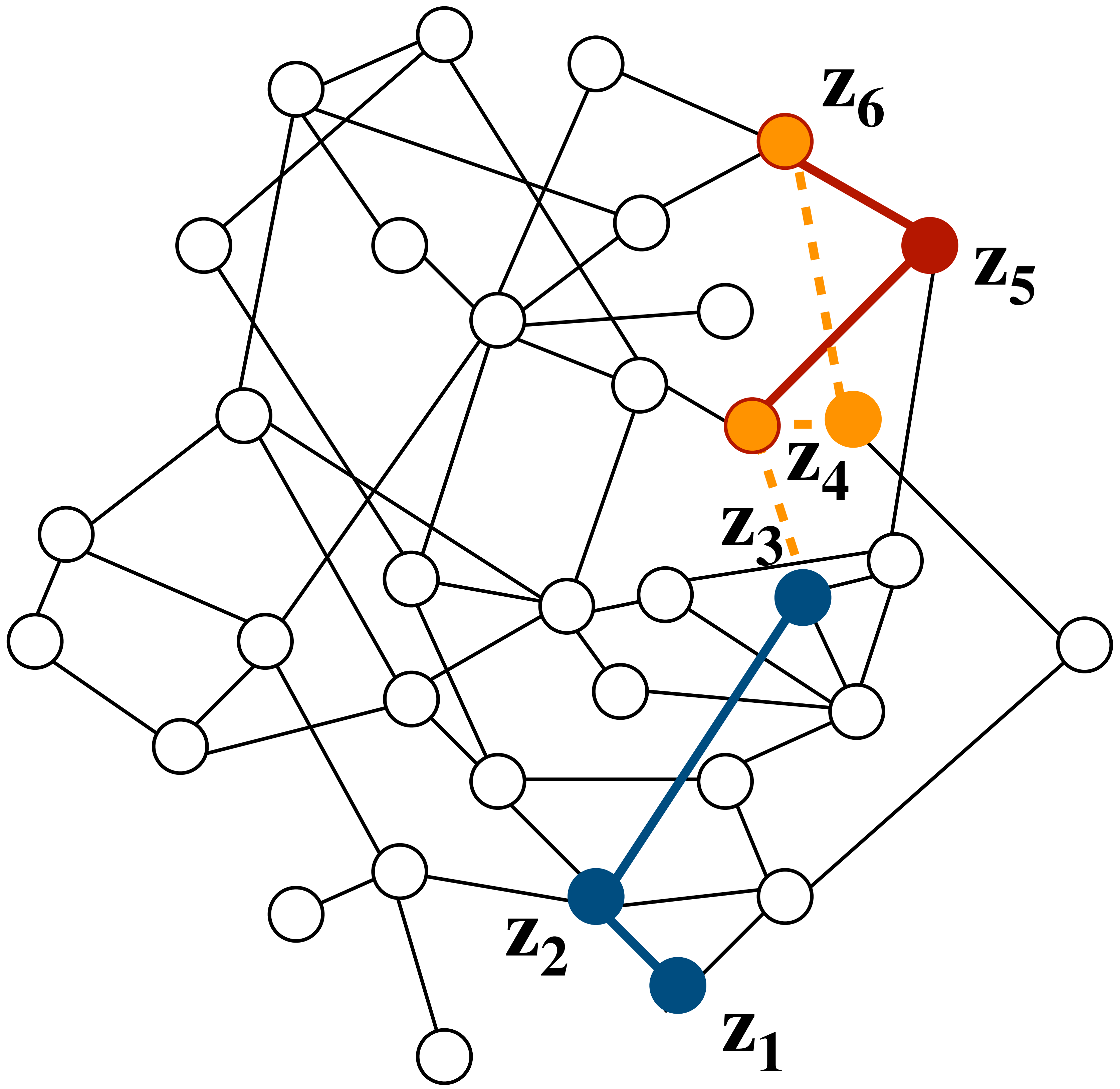
1. Encode **s + r + o** onto random walks

(s, r, o)

(s, r)

2. Train ``reasoning" autoregressive models on **2nd half of random walks** (the half encoding **o**)

reasoning



Hidden Schema Networks for commonsense reasoning

Subject	Relation	Object
PersonX makes PersonY's coffee	xIntent	PersonX wanted to be helpful

TASK: given **Subject + Relation** generate **Object**

Reasoning with **Hidden Schemata**

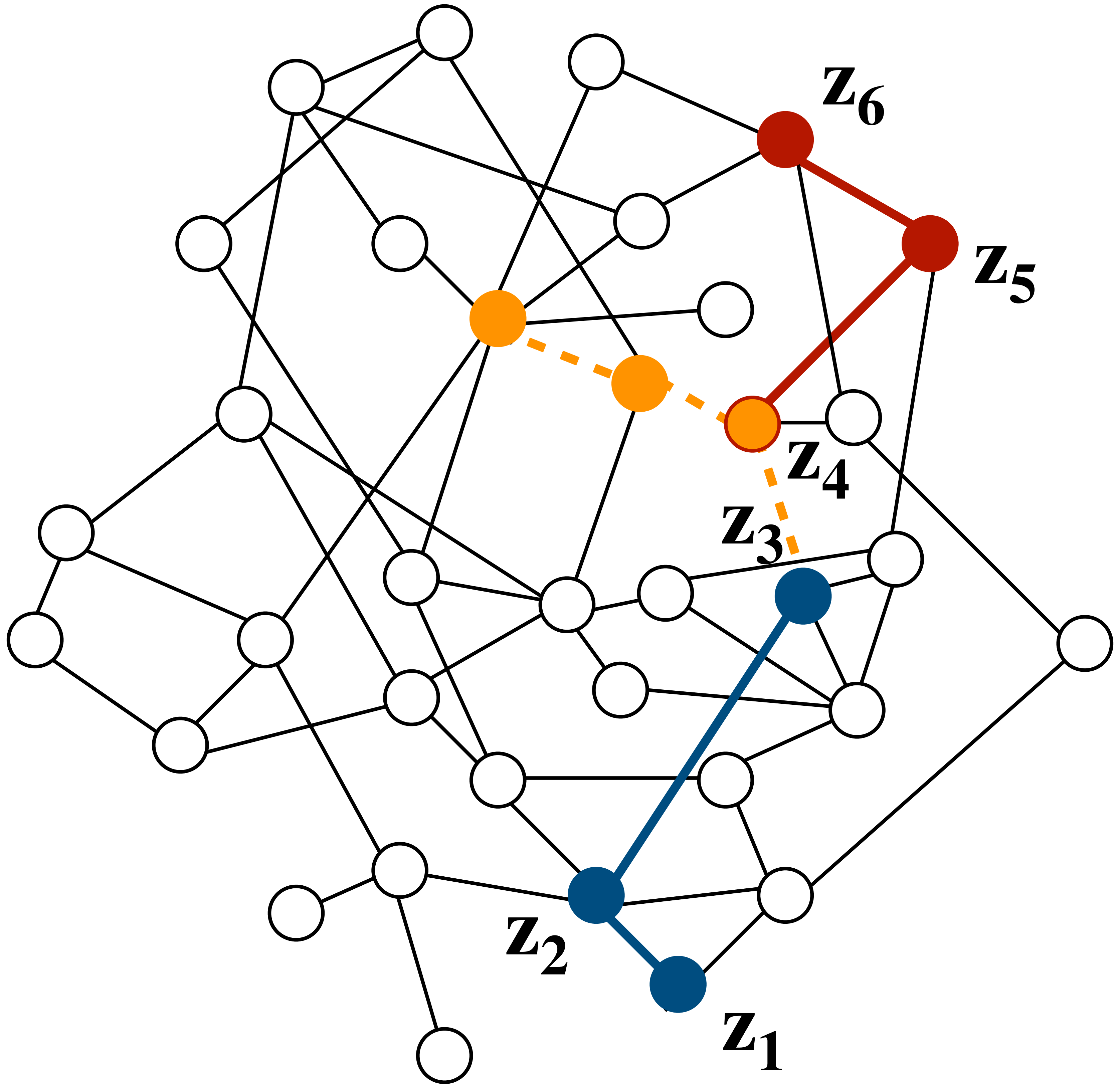
1. Encode **s + r + o** onto random walks

(s, r, o)

(s, r)

2. Train ``reasoning" autoregressive models on **2nd half of random walks** (the half encoding **o**)

reasoning



Hidden Schema Networks for commonsense reasoning

Subject	Relation	Object
PersonX makes PersonY's coffee	xIntent	PersonX wanted to be helpful

TASK: given **Subject + Relation** generate **Object**

Reasoning with **Hidden Schemata**

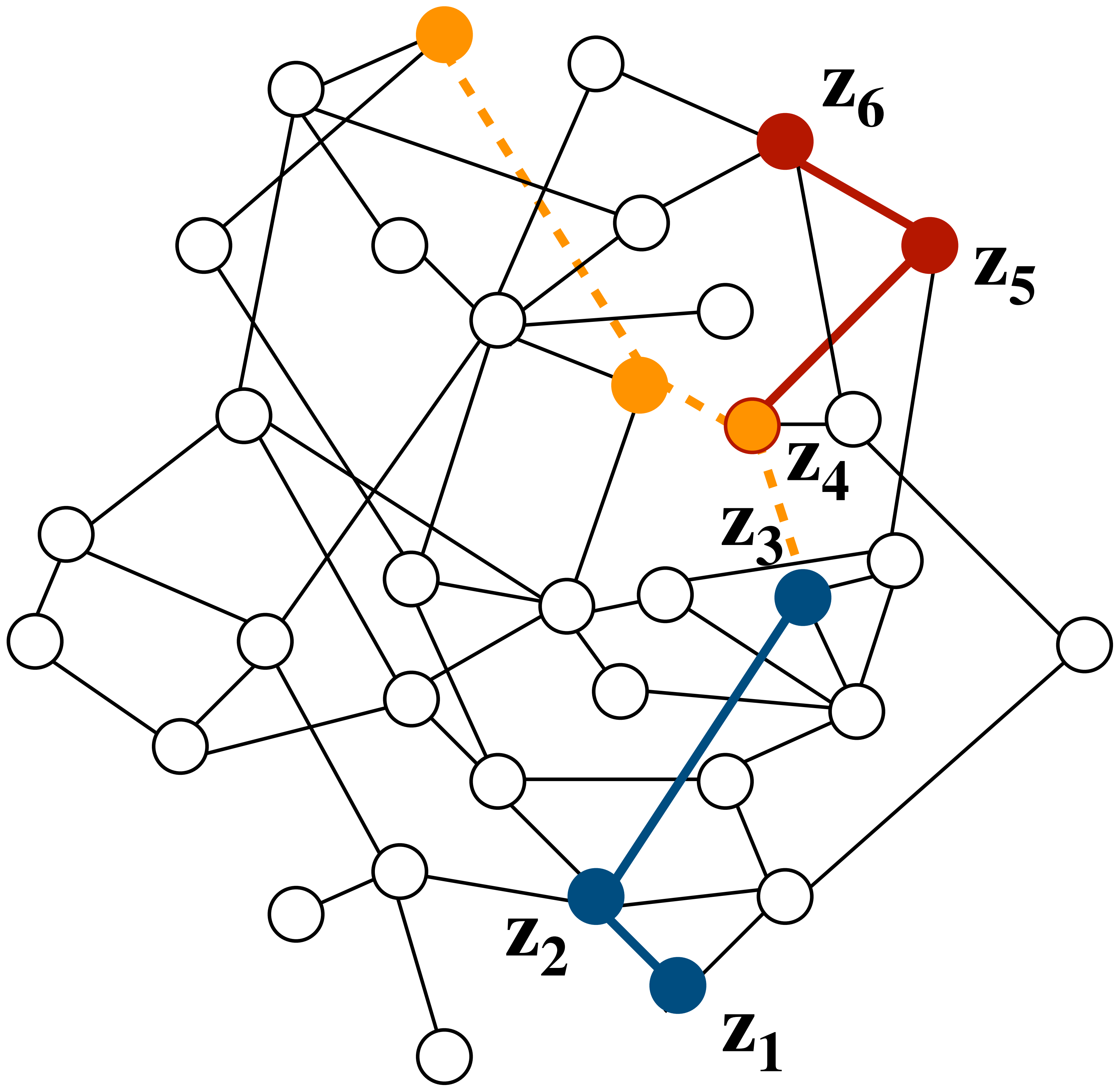
1. Encode **s + r + o** onto random walks

(s, r, o)

(s, r)

2. Train ``reasoning" autoregressive models on **2nd half of random walks** (the half encoding **o**)

reasoning



Hidden Schema Networks for commonsense reasoning

Subject	Relation	Object
PersonX makes PersonY's coffee	xIntent	PersonX wanted to be helpful

TASK: given **Subject + Relation** generate **Object**

Reasoning with **Hidden Schemata**

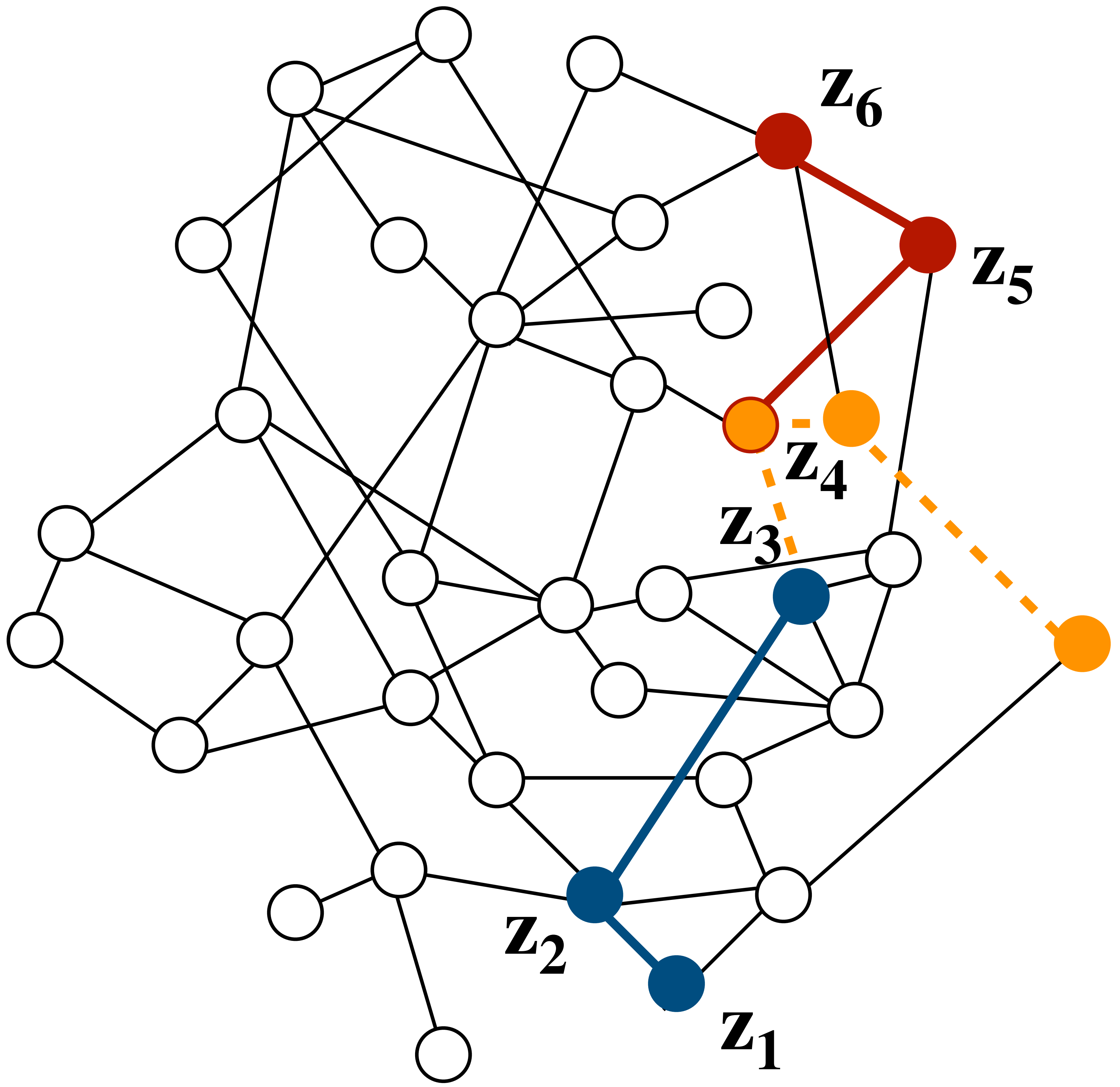
1. Encode **s + r + o** onto random walks

(s, r, o)

(s, r)

2. Train ``reasoning" autoregressive models on **2nd half of random walks** (the half encoding **o**)

reasoning



Hidden Schema Networks for commonsense reasoning

Subject	Relation	Object
PersonX makes PersonY's coffee	xIntent	PersonX wanted to be helpful

TASK: given **Subject + Relation** generate **Object**

Reasoning with **Hidden Schemata**

1. Encode **s + r + o** onto random walks

(s, r, o)

(s, r)

2. Train ``reasoning" autoregressive models on **2nd half of random walks** (the half encoding **o**)

reasoning